STATISTICAL METHODS FOR LINKING THE CHEMICAL COMPOSITION OF PARTICULATE MATTER TO HEALTH OUTCOMES

by

Jenna R. Krall

A dissertation submitted to The Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

March, 2014

Abstract

Short-term exposure to particulate matter air pollution less than 2.5 micrometers in aerodynamic diameter ($PM_{2.5}$) has been associated with mortality and morbidity in epidemiologic studies. $PM_{2.5}$ is a complex mixture of many different chemicals that are emitted from both natural and anthropogenic sources. Recent studies have found that the toxicity of $PM_{2.5}$ depends on its chemical composition, however estimating ambient $PM_{2.5}$ constituent concentrations is challenging in part for several reasons. First, the network of ambient monitors that measure $PM_{2.5}$ chemical constituents is spatially sparse and many communities have only one monitor. Additionally, some constituents of $PM_{2.5}$ are spatially heterogeneous and their concentrations observed at ambient monitors may not be representative of surrounding areas. Last, $PM_{2.5}$ constituents that contribute minimally to $PM_{2.5}$ total mass frequently have censored concentrations that fall below minimum detection limits. Together, these attributes of the available data make estimating ambient $PM_{2.5}$ constituent concentrations difficult and complicate subsequent health effects analyses. We could also estimate health effects of $PM_{2.5}$ sources, which emit combinations of chemical constituents. $PM_{2.5}$ sources are not directly measured and are frequently inferred from $PM_{2.5}$ constituent concentrations. The challenge of estimating $PM_{2.5}$ sources is magnified by the measurement error and data limitations of observed $PM_{2.5}$ constituent concentrations. This dissertation developed methods to address measurement issues in estimating ambient concentrations of $PM_{2.5}$ constituents and $PM_{2.5}$ sources. Then, these methods were used to estimate associations between mortality and short-term exposure to $PM_{2.5}$ constituents and $PM_{2.5}$ sources.

Advisor: Roger D. Peng, Ph.D.

Thesis Readers: Patrick N. Breysse, Ph.D.

Brian S. Caffo, Ph.D. and

Elizabeth C. Matsui, M.D., M.H.S.

Acknowledgments

First, I would like to thank Roger Peng for helping me transition from a first year graduate student to a PhD, and for guiding my development as a researcher. In addition, Roger fostered an incredible research group within the Department of Biostatistics where I received both friendship and mentorship from Brooke Anderson, Amber Hackstadt, and Helen Powell. Through Roger, I also had the opportunity to learn from Michelle Bell and Francesca Dominici, who provided mentorship and vital feedback on Chapters 3 and 4.

Patrick Breysse, Elizabeth Matsui, and Brian Caffo for serving on my thesis committee as well as the very thoughtful questions during my defense and helpful suggestions for my thesis.

Karen Bandeen-Roche for all of her professional and personal support as well as her invaluable example of leadership in both the Department of Biostatistics and the Center on Aging and Health.

Qian-Li Xue for dedicating time to my research interests and for advising me during my training. Also Michelle Carlson for teaching me how to better understand and communicate science.

The faculty, staff, and students in the Department of Biostatistics for providing unending support throughout my time at Johns Hopkins. In particular, my cohort and office mates who helped me through both the academic and personal challenges of completing a PhD.

My parents and brothers, for their love and encouragement and for letting me still call Pittsburgh home. Also my grandparents, for convincing me to come to Hopkins

and for infecting me with a love of knowledge. Last, this would not have been possible without Charles Simpson, who contributed substantially to both the science and computation of Chapter 5, provided crucial editorial advice and feedback, and did not let me subsist on Macaroni and Cheese during grad school.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Particulate matter (PM) is one of six criteria air pollutants regulated by the US Environmental Protection Agency (EPA) through the National Ambient Air Quality Standards. PM air pollution consists of small particles in the air that are generated by both natural and anthropogenic sources. Currently, PM is regulated for two different size distributions: PM less than 2.5 μm in aerodynamic diameter ($PM_{2.5}$) and PM less than 10 μm in aerodynamic diameter (PM_{10}). $PM_{2.5}$ likely represents a more toxic fraction of PM than other size fractions because these smaller particles travel deeper into the lungs, closer to the point of oxygen exchange (Environmental Protection Agency, 2009). Recent epidemiologic studies have found short-term exposure to $PM_{2.5}$ mass is associated with increased hospitalizations and mortality (Burnett et al., 2000; Cifuentes et al., 2000; Peng et al., 2008; Zanobetti and Schwartz, 2009). $PM_{2.5}$ is a spatially and seasonally varying mixture of over 50 chemical constituents (Bell et al., 2007) and it is likely that $PM_{2.5}$ toxicity varies with its chemical composition. Determining the most hazardous $PM_{2.5}$ chemical constituents was identified as a priority by the US National Research Council Committee (National Research Council,

2004). Because $PM_{2.5}$ is a complex mixture of different chemical constituents, regulating $PM_{2.5}$ by composition instead of by total mass may more efficiently protect public health by targeting the most toxic portions of $PM_{2.5}$.

Epidemiologic studies of both fatal and nonfatal health outcomes, including hospitalizations and birth weight, have suggested that health effects vary among individual $PM_{2.5}$ constituents (Ostro et al., 2007; Peng et al., 2009; Zhou et al., 2011; Bell et al., 2010). To estimate health effects of $PM_{2.5}$ chemical constituents in epidemiologic studies, health outcomes such as daily deaths are frequently regressed against daily constituent concentrations. However, the data are generally available at different spatial resolutions with health data aggregated over communities and constituent data observed at ambient monitors, a problem referred to as spatial misalignment. To align aggregated health data and point-level constituent concentrations, researchers frequently estimate the ambient average constituent concentration for a community. The traditional approach for estimating the ambient average is to average observed constituent concentrations from monitors within a community. Estimating ambient $PM_{2.5}$ mass using the traditional approach is likely sufficient because $PM_{2.5}$ is a spatially homogeneous pollutant (Peng et al., 2008; Environmental Protection Agency, 2009) with a large monitoring network. In contrast, the network of monitors that measure $PM_{2.5}$ constituent concentrations is spatially sparse and some $PM_{2.5}$ constituents are spatially heterogeneous (Peng and Bell, 2010), and therefore the traditional approach may not be a good estimate of the ambient average. A spatial model for estimating the ambient average can be used to borrow information from monitors outside the community. Spatial modeling may be better than the traditional approach for estimation when the pollutant is spatially heterogeneous or the monitoring network is

spatially sparse. Using the traditional approach, we first estimated associations between all-cause mortality (excluding accidental deaths) and short-term exposure to 7 major $PM_{2.5}$ constituents, including organic carbon matter (OCM), elemental carbon (EC), silicon, sodium ion, sulfate, nitrate, and ammonium (Chapter 3). Then, we compared estimated mortality effects between ambient $PM_{2.5}$ constituents estimated using the traditional approach and those estimated using a spatial model (Chapter 4).

Estimated associations between $PM_{2.5}$ and mortality vary spatially and seasonally (Zanobetti and Schwartz, 2009), which may correspond with how the composition of $PM_{2.5}$ varies spatially and seasonally (Bell et al., 2007). While observed variation in $PM_{2.5}$ health effects may be driven by variation in chemical composition, observed variation in estimated $PM_{2.5}$ health effects may also result from seasonal or regional differences in human activity patterns, meteorological conditions, penetration of $PM_{2.5}$ indoors, $PM_{2.5}$ sources, or other confounders (Peng et al., 2005). If variation in $PM_{2.5}$ estimated health effects is driven entirely by regional or seasonal variation in $PM_{2.5}$ composition, we would not expect estimated health effects of $PM_{2.5}$ chemical constituents to vary spatially or seasonally. To investigate whether spatial and temporal differences in the estimated health effects of $PM_{2.5}$ are driven by differences in $PM_{2.5}$ composition, we estimated associations between mortality and short-term exposure to $PM_{2.5}$ constituents both nationally, and separately by season and region in the US (Chapter 3).

Another approach to estimating health effects associated with the chemical composition of $PM_{2.5}$ is to estimate health effects of $PM_{2.5}$ sources, which emit combinations of $PM_{2.5}$ chemical constituents. Because chemical constituents generally exist in the air in compounds, estimating health effects of individual $PM_{2.5}$ constituents may identify toxic constituents as well as those correlated with toxic constituents.

Instead, we can target groups of correlated constituents by focusing on estimating health effects of $PM_{2.5}$ sources, which can also inform regulation of certain sources of $PM_{2.5}$. Estimating health effects of $PM_{2.5}$ sources is difficult because concentrations of $PM_{2.5}$ from different sources are not generally available. Frequently, $PM_{2.5}$ sources must be estimated from $PM_{2.5}$ constituent concentrations using source apportionment modeling. Estimation of $PM_{2.5}$ sources depends on observed concentrations of $PM_{2.5}$ constituents, some of which contribute minimally to total mass $PM_{2.5}$ and are frequently censored below minimum detection limits (MDL) (Polissar et al., 2001; Kim et al., 2003; Maykut et al., 2003). Most source apportionment models cannot handle censored data and censored $PM_{2.5}$ constituent concentrations must be imputed or removed before sources are estimated. We compared source estimation between commonly applied methods for handling censored $PM_{2.5}$ constituent data and a new likelihood-based imputation approach (Chapter 5).

Because common source apportionment models can only handle data from one location, most studies estimate health effects of $PM_{2.5}$ sources for one or several communities. In order to combine source apportionment results from multiple monitors, researchers rely on ad hoc approaches to match sources observed at one monitor to sources observed at other monitors. The presence of $PM_{2.5}$ sources can vary between monitors and the same source of $PM_{2.5}$ may have a different estimated chemical composition depending on the location of the monitor (Ito et al., 2004). We developed a method to pool information about $PM_{2.5}$ sources, such as estimated health effects, across multiple monitors. We first determined the major sources of $PM_{2.5}$ shared across multiple ambient monitors and then used these major sources to determine which monitors observed particular sources. In general, our method links sources of $PM_{2.5}$ across multiple ambient monitors. We have developed an approach for pooling

information about $PM_{2.5}$ sources across a region, which allows estimation of regional health effects of $PM_{2.5}$ sources (Chapter 6).

This thesis developed methods for estimating health effects of the chemical composition of $PM_{2.5}$ and used these methods to estimate mortality risks associated with short-term exposure to $PM_{2.5}$ constituents and $PM_{2.5}$ sources. First in Chapter 2, we provided some background on $PM_{2.5}$, $PM_{2.5}$ chemical composition, and $PM_{2.5}$-related health effects. In Chapter 3, we estimated national-level mortality effects of major constituents of $PM_{2.5}$ and determined whether associations varied by region or by season. We investigated whether the method used to adjust for spatial misalignment impacts estimated mortality effects of $PM_{2.5}$ chemical constituents in Chapter 4. With the aim of estimating health effects of $PM_{2.5}$ sources, we first determined how estimation of $PM_{2.5}$ sources is impacted by censored $PM_{2.5}$ constituent concentrations in Chapter 5. Last in Chapter 6, we developed a method to pool information about $PM_{2.5}$ sources across multiple monitoring locations and estimated regional associations between mortality and major $PM_{2.5}$ sources.

Chapter 2

Particulate matter air pollution

$PM_{2.5}$ air pollution is a heterogeneous mixture of particles less than 2.5 micrometers (PM$<$2.5 μm) in aerodynamic diameter and is generated by both anthropogenic and natural sources. Sources that emit $PM_{2.5}$ particles directly into the air are referred to as primary sources and include combustion processes from motor vehicles, coal-fired power plants, vegetative burning, aerosolized sea salt, forest fires, or other industrial sources (Rohr and Wyzga, 2012). Secondary $PM_{2.5}$ refers to particles formed from complex photochemical reactions in the atmosphere. Depending on several factors including weather and particle size, $PM_{2.5}$ can remain suspended in the atmosphere from minutes to weeks and can travel up to thousands of kilometers (Environmental Protection Agency, 2009). $PM_{2.5}$ remains suspended until it is removed from the air by dry deposition, such as settling out onto surfaces, or wet deposition, such as scavenging by precipitation. $PM_{2.5}$ mass is measured by the US EPA Air Quality System (AQS), which is a national ambient monitoring network of over 1,400 monitors with 0 to 3 monitors in each metropolitan statistical area. However, the EPA AQS monitors do not measure the chemical composition of $PM_{2.5}$.

2.1 Chemical composition of $PM_{2.5}$

Concentrations of $PM_{2.5}$ chemical constituents are determined at ambient speciation monitors in the US EPA Chemical Speciation Network (CSN), which is a network of approximately 250 monitors including National Air Monitoring Stations (NAMS) and State and Local Air Monitoring Stations (SLAMS). The EPA CSN monitors are sparsely located throughout the US and this network of monitors is much smaller than the EPA AQS, which measures $PM_{2.5}$ mass. Additionally, EPA CSN monitors were not randomly placed throughout the US. For example, half of the NAMS were placed in areas not in attainment of ozone standards, which may be areas with higher $PM_{2.5}$ concentrations as well (Environmental Protection Agency, 1999). At $PM_{2.5}$ speciation monitors, $PM_{2.5}$ mass is collected on filters and techniques such as X-ray fluorescence, spectrometry, and chromatography are applied to measure concentrations from elements (e.g. nickel and vanadium), anions and cations (e.g. nitrate, sulfate, sodium ion), and carbon constituents (Environmental Protection Agency, 2009). Measurement error in these data can be driven by filter mass measurement bias from the amount of humidity in the air during measurements, loss of volatile particles, electrostatic charge, deposition of additional particles on the filter before or after sampling, or other artifacts (Environmental Protection Agency, 1999).

The EPA CSN measures concentrations of over 50 chemical constituents of $PM_{2.5}$, but most chemical constituents individually contribute less than 1% each to $PM_{2.5}$ total mass on average (Bell et al., 2007). These minor constituents include transition metals, such as copper, zinc, vanadium, nickel, and titanium, and non-metals such as selenium, bromine, and phosphorus. Seven major constituents of $PM_{2.5}$ make up 79-85% of $PM_{2.5}$ total mass on average, both nationally and within the eastern and

western US, including silicon, sulfate, nitrate, ammonium, sodium ion, elemental carbon (EC), and organic carbon matter (OCM). $PM_{2.5}$ total mass is highly correlated with $PM_{2.5}$ ammonium, OCM, nitrate, and sulfate. Other constituents, such as bromine, also covary with $PM_{2.5}$ total mass, but contribute less to total $PM_{2.5}$ by mass (Bell et al., 2007).

The seven major constituents of $PM_{2.5}$ are generated by different sources. EC is primarily generated by combustion processes and is sometimes referred to as black carbon or soot because of its light-absorbing properties (Environmental Protection Agency, 1999). Organic carbon is a mixture of different carbon compounds and can be generated by several sources of $PM_{2.5}$, including combustion processes such as vehicular traffic and vegetative burning (Environmental Protection Agency, 1999). Major secondary inorganic particles found in $PM_{2.5}$ include ammonium, nitrate, and sulfate (Ito et al., 2004; Maykut et al., 2003; Sarnat et al., 2008). Sulfate and nitrate are formed mostly from oxidation of sulfur dioxide and nitrogen dioxide respectively (Schlesinger, 2007). Sulfur dioxide and nitrogen dioxide are generated by fossil fuel combustion, though sulfur dioxide is also generated by natural sources such as oceans and volcanoes (Schlesinger, 2007). Both nitrate and sulfate combine with ammonia in the air to form ammonium sulfate and ammonium nitrate. Therefore, $PM_{2.5}$ ammonium is highly correlated with both $PM_{2.5}$ nitrate and $PM_{2.5}$ sulfate (Bell et al., 2007). Sodium ion, along with chlorine, is frequently generated by salt-related sources, such as aerosolized sea salt. $PM_{2.5}$ from road or soil dust frequently includes crustal material such as silicon and is generated either by wind-blown dust or from mechanical crushing of surfaces, such as cars driving on dirt roads.

Chemical constituents are frequently temporally correlated with one another because they are generated by the same sources. While silicon is generated by road

dust sources, so are other crustal elements such as aluminum, titanium, iron, and calcium (Schlesinger, 2007; Thurston et al., 2011; Maykut et al., 2003; Nikolov et al., 2007; Ito et al., 2004). Aerosolized sea salt primarily consists of sodium and chlorine (Maykut et al., 2003; Thurston et al., 2011), though may also include calcium, magnesium, potassium, and sulfate (Schlesinger, 2007). EC and OCM are found in traffic-related sources of $PM_{2.5}$ (Han et al., 2011; deCastro et al., 2008; Nikolov et al., 2007; Ito et al., 2004) along with potassium (Ito et al., 2004; Sarnat et al., 2008), zinc (Sarnat et al., 2008), nitrate, or copper (Thurston et al., 2011). Coal-fired power plants release sulfur, sulfate, bromine, and selenium (Nikolov et al., 2007; Sarnat et al., 2008) and residual oil sources primarily include nickel and vanadium (Bell et al., 2013; Ito et al., 2004; Thurston et al., 2011). Biomass or vegetative burning releases large amounts of both potassium and carbon (Thurston et al., 2011). The presence of chemical constituents can be driven by increased sources as well as meteorological conditions (deCastro et al., 2008) and concentrations of $PM_{2.5}$ sources and $PM_{2.5}$ chemical constituents vary temporally, even over short periods (Han et al., 2011).

To roughly determine the presence of $PM_{2.5}$ sources across the US, we can compare the sources identified in regional and single-city studies throughout the US. Studies across the US have identified road dust sources (Ito et al., 2004; Hopke et al., 2006; Sarnat et al., 2008; Bell et al., 2013; Nikolov et al., 2007; Lee et al., 2008; Rizzo and Scheff, 2007; Buzcu-Guven et al., 2007; Hwang and Hopke, 2007; Song et al., 2001; Maykut et al., 2003; Larson et al., 2004; Thurston et al., 2011), regional sources (Ito et al., 2004; Hopke et al., 2006; Sarnat et al., 2008; Lee et al., 2008; Rizzo and Scheff, 2007; Buzcu-Guven et al., 2007; Hwang and Hopke, 2007; Song et al.,

2001; Maykut et al., 2003; Larson et al., 2004), including secondary sulfate and nitrate, salt-related sources (Hopke et al., 2006; Bell et al., 2013; Thurston et al., 2011; Rizzo and Scheff, 2007; Hwang and Hopke, 2007; Song et al., 2001; Maykut et al., 2003; Larson et al., 2004) and vegetative burning or wood smoke sources (Hopke et al., 2006; Sarnat et al., 2008; Thurston et al., 2011; Lee et al., 2008; Rizzo and Scheff, 2007; Buzcu-Guven et al., 2007; Hwang and Hopke, 2007; Song et al., 2001; Maykut et al., 2003; Larson et al., 2004). A few studies were able to separately identify traffic sources of $PM_{2.5}$ from motor vehicles and diesel exhaust (Hopke et al., 2006; Sarnat et al., 2008; Hwang and Hopke, 2007; Maykut et al., 2003; Larson et al., 2004), while most other studies across the US identified either a motor vehicle or a diesel source, or simply identified a general "traffic" source that may be a mixture of mobile sources (Ito et al., 2004; Hopke et al., 2006; Bell et al., 2013; Nikolov et al., 2007; Thurston et al., 2011; Lee et al., 2008; Rizzo and Scheff, 2007; Song et al., 2001).

Other sources of $PM_{2.5}$ were more regionally identified. Residual oil or oil incineration sources were primarily found in the northeastern and northwestern US (Ito et al., 2004; Hopke et al., 2006; Bell et al., 2013; Nikolov et al., 2007; Thurston et al., 2011; Song et al., 2001; Maykut et al., 2003; Larson et al., 2004). Power plant-based sources of $PM_{2.5}$ have been identified in Atlanta, GA and Boston, MA (Sarnat et al., 2008; Nikolov et al., 2007). Other sources of $PM_{2.5}$ were identified inconsistently across studies including fireworks, coal combustion, lead smelter, railroad, steel industry and sources with one tracer element such as selenium, iron, copper, and manganese (Hopke et al., 2006; Thurston et al., 2011; Sarnat et al., 2008; Lee et al., 2008; Rizzo and Scheff, 2007; Buzcu-Guven et al., 2007).

The chemical composition of $PM_{2.5}$ in the US varies spatially and temporally

(Environmental Protection Agency, 2009; Bell et al., 2007; Han et al., 2011). Higher concentrations of nitrate, chlorine, zinc, nickel, and bromine in the winter (Environmental Protection Agency, 2009; Bell et al., 2007) could be attributable to increased use of oil heating. Road dust constituents such as aluminum, titanium, magnesium, silicon, and sulfate are higher in the summer (Environmental Protection Agency, 2009; Bell et al., 2007), a trend that could be driven by seasonal wind patterns. In addition, there are regional trends in the composition of $PM_{2.5}$. Sulfate $PM_{2.5}$ is higher in the eastern half of the US, while nitrate is higher in the west (Environmental Protection Agency, 2009; Bell et al., 2007). Since sulfate and nitrate are primarily secondary pollutants, regional differences in their concentrations could be attributable to regional differences in the concentrations of their generating gaseous pollutants, sulfur dioxide and nitrogen dioxide, or regional weather differences. Sodium, which frequently is generated by sea salt released into the air, is higher near coastal areas (Bell et al., 2007).

2.2 Health effects of particulate matter

$PM_{2.5}$ can impact different organ systems of the human body including the pulmonary system, circulatory system, and central nervous system (CNS) (Environmental Protection Agency, 2009). There are many hypothesized mechanisms for how exposure to $PM_{2.5}$ leads to adverse health outcomes, including pulmonary or systemic inflammation, oxidative stress, and altered cardiac autonomic function (Pope and Dockery, 2006; Pope et al., 2004). $PM_{2.5}$ reacts with cells in the lungs to form reactive oxygen species (ROS), which signal the release of proinflammatory molecules. Increased pulmonary inflammation can exacerbate chronic obstructive pulmonary disease (COPD) and asthma (Environmental Protection Agency, 2009). In particular,

$PM_{2.5}$ containing metals can drive the creation of ROS that leads to oxidative stress and increased immune response (Environmental Protection Agency, 2009; Schwarze et al., 2006; Lippmann et al., 2006; Nel, 2005).

Chemical constituents that contribute minimally to $PM_{2.5}$ by mass, such as metals, organic substances, and soluble $PM_{2.5}$ constituents (e.g. zinc and copper), can translocate directly from the lungs into the circulatory system (Environmental Protection Agency, 2009; Schwarze et al., 2006). Once in the circulatory system, these particles can cause inflammation in the heart or affect cardiac autonomic function. Inflammatory markers that enter into the blood stream from the pulmonary system can also alter the coagulation of the blood and lead to myocardial infarction (Schwarze et al., 2006). Other cardiovascular outcomes related to $PM_{2.5}$ exposure include atherosclerosis and decreases in heart rate variability (Grahame and Schlesinger, 2010). Effects of exposure to $PM_{2.5}$ on the CNS can be driven by constituents translocating into the olfactory bulb and into the CNS (Schlesinger, 2007).

The evidence for adverse health effects of $PM_{2.5}$ has been collected through concentrated air particles (CAPs) studies, animal studies, in-vitro studies of human cells, and epidemiologic studies of human health. In a rat model for chronic bronchitis, Batalha et al. (2002) found increased vasoconstriction was associated with CAPs exposure. Brown et al. (2001) instilled particles directly into the lungs of rats and found increased inflammatory response for smaller particles. Reduced heart rate variability and increased arrhythmias were found in mice exposed to ambient PM (Wang et al., 2012). Veronesi et al. (2002) found that surface charge of PM determines the inflammatory response in human epithelial cells. Human bronchial epithelial cells exposed to PM have shown increased reactive oxygen species production and increased Interleukin-6 (Zhao et al., 2009).

Epidemiologic studies of $PM_{2.5}$ have been critical for determining the impact of ambient $PM_{2.5}$ levels on human health and quantifying the potential public health impact of reducing $PM_{2.5}$ emissions. Short-term exposure to $PM_{2.5}$ has been associated with increased risk of mortality and morbidity in recent epidemiologic studies (Dominici et al., 2006; Mar et al., 2000; Pope et al., 2002; Ostro et al., 2006; Zanobetti and Schwartz, 2009). Specifically large, national-level studies have provided substantial evidence for associations of $PM_{2.5}$ with adverse health outcomes including a study of mortality in 112 US cities (Zanobetti and Schwartz, 2009) and a study of cardiovascular and respiratory hospitalizations in 204 urban US counties (Dominici et al., 2006). Long-term exposure to $PM_{2.5}$ has been associated with mortality and lung cancer (Puett et al., 2009; Pope et al., 2002). The EPA's Integrated Science Assessment classified associations between $PM_{2.5}$ and cardiovascular outcomes as "causal" and associations between $PM_{2.5}$ and respiratory outcomes as "likely to be causal," for both short-term and long-term exposure to $PM_{2.5}$(Environmental Protection Agency, 2009).

2.2.1 Health effects of the chemical composition of $PM_{2.5}$

Literature reviews of the health effects of different $PM_{2.5}$ constituents and $PM_{2.5}$ sources have been conducted by Schlesinger (2007), Rohr and Wyzga (2012), and Grahame and Schlesinger (2007). Here, we reviewed some of the relevant results from the literature supporting the toxicity of short-term exposure to $PM_{2.5}$ constituents and sources of $PM_{2.5}$.

Toxicological evidence

Evidence from toxicological studies have found different $PM_{2.5}$ constituents to be most associated with health. Constituents of $PM_{2.5}$ associated with road dust or soil, such as aluminum and silicon, have been found to be harmful in toxicological studies (Schlesinger, 2007). In animal CAPs studies, increased exposure to silicon was associated with vasoconstriction and increased myocardial ischemia heart rate (Batalha et al., 2002; Wellenius et al., 2003). Another CAPs study of canines found myocardial ischemia to be associated with exposure to road dust (Nikolov et al., 2007). Both OC and EC have been associated with health outcomes for both in vitro and in vivo toxicological studies (National Research Council, 2004; Urch et al., 2004). Increased exposure to EC and OCM were associated with decreased changes in brachial artery diameter in a CAPs study of humans (Urch et al., 2004). In rats and mice, increases in EC and OCM were also associated with increases in bronchoalveolar lavage neutrophils and increased allergic inflammatory responses (Godleski et al., 2002; Kleinman et al., 2007). EC and OCM are commonly generated by diesel or traffic sources of PM, and exposure to CAPs associated with motor vehicles was found to increase airway irritation in canines (Nikolov et al., 2008). A study of alveolar epithelial cells also found that exposure to diesel-related PM increased Interleukin-8, a chemokine associated with inflammation (Seagrave et al., 2004). In general, sulfate has not been found to be harmful at ambient levels in toxicological studies (Schlesinger, 2007). While some toxicological studies have estimated health effects related to sodium ion and nitrate exposure, the results were commonly null or mixed (Schlesinger, 2007).

Transition metals and other nonmetal constituents that contribute less to $PM_{2.5}$ by mass have been implicated in toxicological studies, including CAPs and in vitro

studies (Costa and Dreher, 1997; Lippmann et al., 2006; Huang et al., 2003; Gojova et al., 2007). These metals have been found to translocate deep in the lungs of rats (Rohr et al., 2010). Nickel and vanadium, commonly found in residual oil sources of $PM_{2.5}$, have been linked to bradycardia, arrhythmogenesis, and hypothermia in rats (Campen et al., 2002). In human epithelial cells, nickel was also associated with "hypoxia-like" stress (Salnikow et al., 2004). In a rat model of chronic bronchitis, bromine and vanadium were associated with bronchoalveolar lavage neutrophils (Saldiva et al., 2002).

Epidemiologic studies

Like toxicological studies, epidemiologic studies of $PM_{2.5}$ constituents have each identified a different subset of constituents as most harmful to human health (Ito et al., 2011; Peng et al., 2009; Zhou et al., 2011; Ostro et al., 2007). A recent literature review found more epidemiologic evidence supporting the toxicity of OCM and EC than other constituents (Rohr and Wyzga, 2012). In a large study of hospitalizations in the US, OCM and EC were associated with cardiovascular hospital admissions and OCM was associated with respiratory hospital admissions (Peng et al., 2009). Other studies have found associations of OC and EC with cardiovascular hospitalizations and emergency department visits (Ito et al., 2011; Tolbert et al., 2007; Kim et al., 2012), EC with lower birthweight (Bell et al., 2010), and OC and EC with mortality (Ito et al., 2011; Zhou et al., 2011). Zanobetti et al. (2009) found OC and EC modified the association between $PM_{2.5}$ and Diabetes hospitalizations and Bell et al. (2009) found EC modified the association between $PM_{2.5}$ and cardiovascular and respiratory hospitalizations.

$PM_{2.5}$ metals may be more harmful to human health because of their role in creating reactive oxygen species (Schwarze et al., 2006), however $PM_{2.5}$ metals have been analyzed in fewer epidemiologic studies than other constituents. Gestational exposure to nickel, vanadium, zinc, and aluminum has been associated with lower birthweight (Bell et al., 2010). In a national US-based study, larger associations between short-term exposure to $PM_{2.5}$ and hospitalizations were found in communities with higher vanadium and nickel concentrations (Bell et al., 2009). A study of cardiovascular mortality and hospitalizations in New York City found some evidence for associations between hospitalizations and short-term exposure to nickel and zinc, but little evidence of associations with vanadium (Ito et al., 2011). Zhou et al. (2011) found zinc and potassium to be associated with mortality in Seattle, but they did not find vanadium or nickel to be associated with hospitalizations or mortality in Seattle or Detroit. In California, Ostro et al. (2007) found evidence of associations of mortality with copper, potassium, titanium, and zinc, but not vanadium or nickel. A national US-based study found nickel, aluminum, and arsenic modified the association between $PM_{2.5}$ and mortality, but vanadium and zinc did not (Franklin et al., 2008). In a similar national-level study, the association between $PM_{2.5}$ and cardiovascular hospitalizations was found to be modified by nickel, sodium ion, vanadium, and aluminum (Zanobetti et al., 2009). Sodium ion has also been associated with cardiovascular disease mortality (Ito et al., 2011).

Other constituents have also been found to be associated with adverse health outcomes. Silicon has been associated with decreased birthweight, hospitalizations, and mortality (Bell et al., 2010; Ito et al., 2011; Zhou et al., 2011). Exposure to sulfate has been associated with preterm birth, asthma hospitalizations, and mortality (Darrow et al., 2009; Ito et al., 2011; Kim et al., 2012; Ito et al., 2011; Cao et al.,

2012). Sulfate has also been found to modify the association between $PM_{2.5}$ and hospitalizations (Zanobetti et al., 2009). Nitrate has been associated with asthma and cardiovascular disease hospitalizations as well as mortality (Kim et al., 2012; Peng et al., 2009; Cao et al., 2012; Ostro et al., 2007). Other constituents, such as selenium, bromine, and ammonium have also been associated with adverse health outcomes (Ito et al., 2011; Cao et al., 2012).

Recently, epidemiologic studies have estimated associations between sources of $PM_{2.5}$ and health outcomes. Evidence for the toxicity for $PM_{2.5}$ sources, like the evidence for $PM_{2.5}$ constituents, is not very consistent across studies. In Phoenix, AZ, there was more evidence of an association between mortality and exposure to $PM_{2.5}$ copper smelter, traffic, secondary sulfate, and sea salt and less evidence of an association with biomass burning and soil sources (Mar et al., 2006). A study in Washington, DC found copper and secondary sulfate sources of $PM_{2.5}$ to be most associated with mortality (Ito et al., 2006). $PM_{2.5}$ from mobile sources and coal combustion, but not crustal particles, were associated with daily mortality in six US cities (Laden et al., 2000). The $PM_{2.5}$ sources most associated with cardiovascular emergency department visits in Atlanta, GA were mobile sources, including diesel and gasoline engines, and biomass burning (Sarnat et al., 2008). In four counties in Connecticut and Massachusetts, road dust and sea salt were associated with respiratory hospitalizations (Bell et al., 2013). One study of reproductive outcomes found third trimester exposure to an oil combustion source of $PM_{2.5}$ was associated with lower birthweight (Bell et al., 2010). In general, larger epidemiologic studies of associations between adverse health outcomes and short-term exposure to $PM_{2.5}$ constituents and $PM_{2.5}$ sources are needed to conclusively determine which portions of $PM_{2.5}$ are most toxic.

Chapter 3

Short-term exposure to particulate matter constituents and mortality in a national study of U.S. urban communities

This chapter is reproduced with permission from *Environmental Health Perspectives*, see Krall et al. (2013).

Although the association between $PM_{2.5}$ mass and mortality has been extensively studied, few national-level analyses have estimated mortality effects of $PM_{2.5}$ chemical constituents. Epidemiologic studies have reported that estimated effects of $PM_{2.5}$ on mortality vary spatially and seasonally. We hypothesized that associations between $PM_{2.5}$ constituents and mortality would not vary spatially or seasonally if variation in chemical composition contributes to variation in estimated $PM_{2.5}$ mortality effects. We aimed to provide the first national, season-specific, and region-specific associations between mortality and $PM_{2.5}$ constituents. We estimated short-term associations between nonaccidental mortality and $PM_{2.5}$ constituents across 72 urban U.S. communities from 2000 to 2005. Using U.S. Environmental Protection Agency

(EPA) Chemical Speciation Network data, we analyzed seven constituents that together compose 79-85% of $PM_{2.5}$ mass: organic carbon matter (OCM), elemental carbon (EC), silicon, sodium ion, nitrate, ammonium, and sulfate. We applied Poisson time-series regression models, controlling for time and weather, to estimate mortality effects. Interquartile range increases in OCM, EC, silicon, and sodium ion were associated with estimated increases in mortality of 0.39% [95% posterior interval (PI): 0.08, 0.70%], 0.22% (95% PI: 0.00, 0.44), 0.17% (95% PI: 0.03, 0.30), and 0.16% (95% PI: 0.00, 0.32), respectively, based on single-pollutant models. We did not find evidence that associations between mortality and $PM_{2.5}$ or $PM_{2.5}$ constituents differed by season or region. Our findings indicate that some constituents of $PM_{2.5}$ may be more toxic than others and, therefore, regulating PM total mass alone may not be sufficient to protect human health.

3.1 Introduction

The previously observed associations between $PM_{2.5}$ and mortality and morbidity might be driven by one or several of the chemical constituents of $PM_{2.5}$ that covary most with $PM_{2.5}$ total mass or contribute most to $PM_{2.5}$ by mass. For 72 U.S. urban communities, we estimated national-level associations between mortality and short-term exposure to seven major chemical constituents of $PM_{2.5}$: OCM, EC, silicon, sodium ion, nitrate, ammonium, and sulfate. In order to determine whether previously observed spatial and temporal variations in the health effects of $PM_{2.5}$ were driven by spatial and temporal variations in the chemical composition of $PM_{2.5}$, we also estimated mortality effects of $PM_{2.5}$ constituents by season and by region across the US. To our knowledge, this is the first national-level U.S. study to estimate the effects of short-term exposure to individual $PM_{2.5}$ constituents on human mortality.

3.2 Methods

3.2.1 Mortality data

All-cause mortality data (excluding accidental deaths) were aggregated from death certificate data obtained from the National Center for Health Statistics for 2000 to 2005 (Samet et al., 2000b). The original database includes mortality data for 108 urban communities (each consisting of one county or set of adjacent counties). For the present analysis, we excluded communities that were located outside the continental United States (n = 2 communities) or that had no $PM_{2.5}$ constituent monitors (n = 29), no days with data for all seven $PM_{2.5}$ constituents during 2000-2005 (n = 4), or insufficient data for model convergence (n = 1), leaving 72 communities for our analysis.

3.2.2 $PM_{2.5}$ constituent and weather data

We obtained $PM_{2.5}$ constituent data for 2000-2005 from the U.S. Environmental Protection Agency (EPA) Chemical Speciation Network, which records concentrations of >50 chemical constituents that contribute to $PM_{2.5}$ mass from approximately 250 monitoring sites throughout the continental United States (Bell et al., 2007; Peng et al., 2009). For daily concentrations of $PM_{2.5}$ mass, we used data from the U.S. EPA Air Quality System from 2000 to 2005, which included approximately 1,400 monitoring sites (Dominici et al., 2006; Peng et al., 2009). We excluded data from source-oriented monitors that may not be representative of typical population exposures.

We analyzed a subset of seven constituents previously identified as covarying with $PM_{2.5}$ total mass and/or having the largest contribution to overall $PM_{2.5}$ total mass: OCM, EC, silicon, sodium ion, nitrate, ammonium, and sulfate (Bell et al., 2007). Together, these constituents account for 79-85% of yearly and seasonal $PM_{2.5}$ mass (both nationally and in the eastern and western United States). Other constituents each contribute <1% on average to the total $PM_{2.5}$ mass (Bell et al., 2007).

Monitors typically measure $PM_{2.5}$ constituent concentrations every third or sixth day. Organic carbon measurements were adjusted for field blanks to estimate OCM using a standard approach such that OCM = $1.4(OC_m - OC_b)$, where OC_m represents measured organic carbon, OC_b represents organic carbon for blank filters, and 1.4 is the adjustment factor to account for non-carbon organic matter, as described previously (Bell et al., 2007).

We estimated daily community-level pollutant exposure as the arithmetic mean of daily monitor observations within the community. For communities with a single

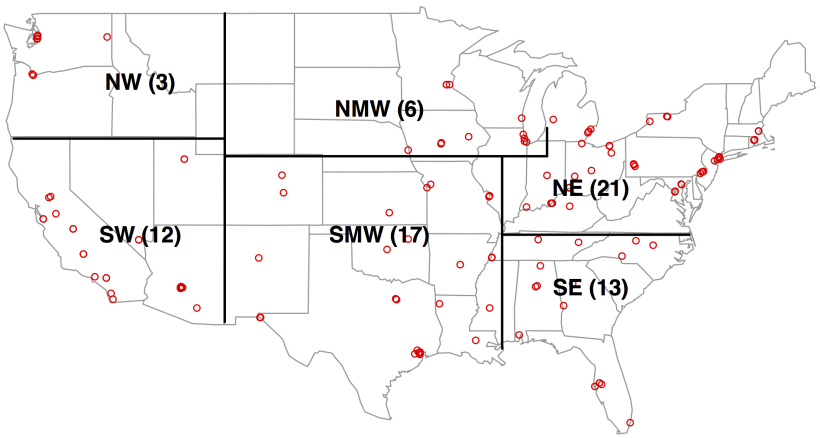

Figure 3.1: Map of the United States illustrating the 72 U.S. communities analyzed (red circles) divided into the six regions used in this analysis: NE, northeast; NMW, north midwest; NW, northwest; SE, southeast; SMW, south midwest; SW, southwest. Numbers in parentheses indicate the number of study communities within that region.

monitor, we used pollutant concentrations recorded by that monitor.

We divided the United States into six regions based loosely on U.S. EPA regions (Figure 3.1). Similar divisions have been used in other studies to approximately reflect variation in $PM_{2.5}$ sources (Peng et al., 2005; Samet et al., 2000b; Zanobetti and Schwartz, 2009). Daily temperature and dew point temperature were obtained from the National Oceanic and Atmospheric Administration (EarthInfo Inc., 2006; Peng et al., 2009).

3.2.3 Mortality risk model

We modeled short-term associations between mortality counts and $PM_{2.5}$ constituent concentrations with overdispersed log-linear Poisson time-series regression models. For each constituent considered, we fit a separate community-specific single-pollutant model. We chose additional covariates based on previous analyses (Peng et al., 2009; Zanobetti and Schwartz, 2009). These covariates included smooth functions (natural spline) of temperature [degrees of freedom (df) = 3], 1-day lag of temperature (df = 3), and long-term and seasonal trends in mortality (df = 8/year) as well as categorical variables for age ($<$65, 65-74, $>$74 years) and day of week. We also estimated associations between $PM_{2.5}$ mass and mortality.

Past research identified previous-day $PM_{2.5}$ exposure as the exposure lag most strongly associated with mortality (Ito et al., 2011; Samet et al., 2000b), and studies of $PM_{2.5}$ constituents have corroborated this finding (Huang et al., 2012; Ito et al., 2011). We therefore included the mean value of each pollutant on the previous-day (lag 1) in single-pollutant mortality risk models. As a sensitivity analysis, we estimated mortality effects of mean exposure on the same day (lag 0) and 2 days before (lag 2). Because constituent data were not collected on consecutive days, we could not estimate effects using distributed lag models (Dominici et al., 2006).

We estimated season-specific effects by adding interaction terms between pollutant concentration and seasons to our mortality risk model. The four seasons were winter (21 December-20 March), spring (21 March-20 June), summer (21 June-20 September), and fall (21 September-20 December) (Peng et al., 2005).

To estimate national, seasonal, and regional mortality effects, we combined community -specific mortality risk estimates using a two-level normal Bayesian hierarchical model (Peng et al., 2009). To facilitate comparisons across pollutants, we report results as percent increases in mortality risk for an interquartile range (IQR) increase in pollutant concentration, with corresponding 95% Bayesian posterior intervals (95% PIs). We also report posterior probabilities that the mortality risk associated with a pollutant is greater than 0 ($p > 0$).

To analyze differences in estimated pollutant effects by season, we pooled the community-specific estimated mortality risk differences comparing each season to winter in order to obtain national-level 95% PIs for the seasonal differences. We concluded that there was no evidence of seasonal differences if these posterior intervals included zero. Because we fit separate time-series models for each community in the study, we were unable to use this same approach to explore regional differences in mortality risk. To analyze differences in risks by region, we used the pooled region-specific estimates and estimated 95% PIs for pairwise differences in mortality effect estimates between regions.

3.3 Results

3.3.1 Summary statistics

Study communities had a combined population of 88.4 million people (2000 census) (US Census Bureau, 2013), with 0-254 daily nonaccidental deaths (median, 15 deaths/day). For each pollutant, the mean, minimum, and maximum days of data used in community-specific models are shown in Table 3.1. Although data were limited by the nondaily sampling schedule of $PM_{2.5}$ constituent monitors, most communities (67 of 72) had $\geq$ 150 days of constituent data. We restricted the constituent

Table 3.1: Mean (minimum-maximum) number of days of observation in the study period used for community-specific mortality risk models, IQR (median of monitor-specific IQRs), and median (minimum-maximum) community-specific average constituent concentration ($\mu g/m^3$).

Pollutant	No. of days	IQR	Concentration
$PM_{2.5}$	1,636 (456 - 2,189)	8.00	13.6 (6.38 - 22.84)
OCM	388 (58 - 907)	3.08	4.15 (2.22 - 8.89)
EC	395 (58 - 921)	0.37	0.68 (0.29 - 1.51)
Silicon	395 (56 - 920)	0.08	0.11 (0.05 - 0.52)
Sodium ion	374 (58 - 834)	0.11	0.12 (0.04 - 0.60)
Nitrate	387 (58 - 720)	1.22	1.70 (0.50 - 10.05)
Ammonium	392 (58 - 923)	1.14	1.53 (0.34 - 3.90)
Sulfate	392 (58 - 923)	2.75	3.50 (0.71 - 5.91)

Table 3.2: Pairwise correlations for $PM_{2.5}$ chemical constituents for all seasons obtained by taking the median of all monitor location-specific correlations.

	EC	Silicon	Sodium ion	Nitrate	Ammonium	Sulfate
OCM	0.64	0.20	0.10	0.22	0.47	0.42
EC	1.00	0.10	0.04	0.33	0.34	0.19
Silicon		1.00	0.09	-0.07	0.05	0.15
Sodium			1.00	0.12	0.04	0.10
Nitrate				1.00	0.56	0.08
Ammonium					1.00	0.87
Sulfate						1.00

monitor data to monitors located within the community boundaries (n = 141). Most communities had only one monitor collecting data (n = 39 communities). The other 33 communities had two monitors (n = 18 communities), three monitors (n = 9), five monitors (n = 2), seven monitors (n = 3), or eight monitors [n = 1 (New York City)]. Across communities, median concentrations of sulfate and OCM tended to be higher than other $PM_{2.5}$ constituents (Table 3.1). Within communities, sulfate and ammonium, and OCM and EC, were highly correlated (correlation coefficients of 0.87 and 0.64, respectively); otherwise, correlations between constituent pairs were moderate or weak (Table 3.2).

Table 3.3: National average estimated percent increase (95% PI) in mortality associated with an IQR increase in $PM_{2.5}$ constituents on the previous day for single-pollutant and multipollutant models.

	Single pollutant models		Multipollutant model[a]	
Pollutant	Estimate (95% PI)	PP(>0)	Estimate (95% PI)	PP(>0)
$PM_{2.5}$	0.30 (0.11,0.50)	1.00		
OCM	0.39 (0.08,0.70)	0.99	0.23 (-0.46,0.92)	0.74
EC	0.22 (0.00,0.44)	0.97	0.14 (-0.38,0.65)	0.70
Silicon	0.17 (0.03,0.30)	0.99	0.19 (0.00,0.38)	0.97
Sodium ion	0.16 (0.00,0.32)	0.98	0.10 (-0.23,0.44)	0.72
Nitrate	0.07 (-0.10,0.24)	0.80		
Ammonium	0.02 (-0.25,0.29)	0.56		
Sulfate	-0.02 (-0.38,0.35)	0.46		

PP, posterior probability.

[a] Explores whether the associations between mortality OCM, EC, silicon, and sodium ion in single-pollutant models are confounded by a subset of these four constituents.

3.3.2 Mortality risk estimates

We estimated that mortality increased by 0.39% (95% PI: 0.08, 0.70) in association with an IQR increase in OCM on the previous day. Mortality was also associated with IQR increases in EC (0.22%; 95% PI: 0.00, 0.44), silicon (0.17%; 95% PI: 0.03, 0.30), and sodium ion (0.16%; 95% PI: 0.00, 0.32) (Table 3.3, Figure 3.2). The posterior probability of a positive association with mortality for each of these constituents was > 0.95.

We also estimated season-specific (Figure 3.3) and region-specific (Figure 3.4) mortality effects of $PM_{2.5}$ constituents. We found evidence of a season-specific effect of an IQR increase in silicon on the previous day during the summer (0.23%; 95% PI: 0.03, 0.44), but no other season-specific or region-specific effect estimates were statistically significant, and we found no evidence that estimated effects of any of the seven $PM_{2.5}$ constituents varied by season or by region using posterior intervals of

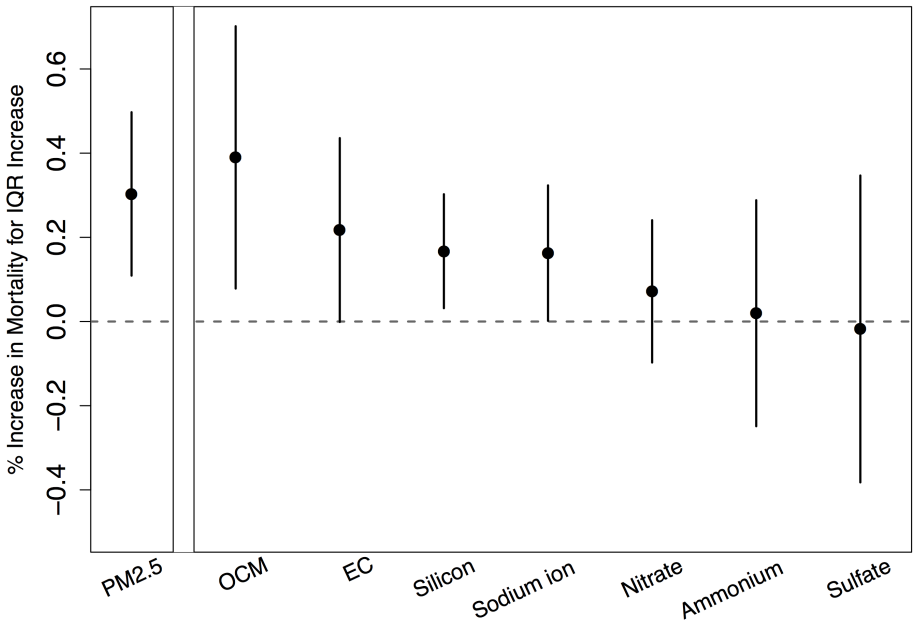

Figure 3.2: National average estimated percent increase in mortality (95% PI) associated with an IQR increase in $PM_{2.5}$ constituents on the previous day for single-pollutant models.

the differences across seasons and regions.

An IQR increase in $PM_{2.5}$ mass on the previous day (8.00 $\mu g/m^3$) was associated with a 0.30% increase in mortality (95% PI: 0.11, 0.50) (Table 3.3, Figure 3.2). We found no evidence that associations between $PM_{2.5}$ mass and mortality varied strongly by season or region, although some season-specific and region-specific associations between $PM_{2.5}$ and mortality did rise to the level of statistical significance. For example, a 0.37% increase in mortality (95% PI: 0.05, 0.69) was associated with an IQR increase in $PM_{2.5}$ in the northeast region (Figure 3.4).

As a sensitivity analysis, we estimated same-day and 2-day lagged national-average, season-specific, and region-specific mortality risks associated with $PM_{2.5}$ and $PM_{2.5}$ constituents, although we found little evidence of associations with mortality at these lags (Table 3.4, Figures 3.5-3.8). The national-average associations of same-day sulfate (0.29%; 95% PI: -0.10, 0.68) and ammonium (0.11%; 95% PI: -0.20, 0.42) with mortality were larger in magnitude than previous-day associations (Table 3.4). We found some indication that same-day $PM_{2.5}$ was associated with mortality nationally and in the spring and summer (Table 3.4, Figure 3.5).

3.3.3 Sensitivity analyses

We considered several variations of our primary mortality risk model: adding a linear term for dew point temperature, increasing the degrees of freedom for both smooth functions of temperature, and including different degrees of freedom for the smooth function of time (4, 6, 10, or 12 df/year). We also tested the sensitivity of our seasonal model to season definition (winter: 1 December-28 February), and none of these alternate models produced substantially different mortality risk estimates (results not shown). When we limited data to consider cardiovascular and respiratory mortality,

we found estimated effects similar to all-cause mortality (results not shown).

We fit a multipollutant mortality risk model including OCM, EC, silicon, and sodium ion simultaneously to assess whether associations found for OCM, EC, silicon, and sodium ion in single-pollutant models could be due to confounding by a subset of these four constituents (Table 3.3). Compared with single-pollutant model estimates, multipollutant mortality risk estimates were slightly attenuated for OCM, EC, and sodium ion and slightly increased for silicon, indicating that there was little joint confounding by the four constituent exposures. Multipollutant estimates were based on an average of 358 days of data compared with an average of 389 days for single-pollutant models. Therefore, multipollutant model estimates had larger standard errors and smaller posterior probabilities of being greater than zero than their single-pollutant counterparts.

3.4 Discussion

We conducted a national-level study to estimate national, seasonal, and regional associations between mortality and short-term exposures to seven major constituents of $PM_{2.5}$ mass in 72 U.S. urban communities from 2000 to 2005. Among the seven constituents examined in this study, OCM, EC, silicon, and sodium ion were most strongly associated with mortality, with high posterior probabilities of a mortality risk larger than zero in single-pollutant models of exposure on the previous day. Epidemiologic, toxicological, and controlled human exposure studies have reported associations of EC and OCM with adverse health outcomes (Ito et al., 2011; Ostro et al., 2007; Peng et al., 2009; Rohr and Wyzga, 2012; Tolbert et al., 2007). In a literature review, Rohr and Wyzga (2012) concluded that evidence supporting the toxicity of

carbon-containing constituents might be stronger than for other constituents. Previous work has also indicated that silicon may be more toxic than other constituents (Franklin et al., 2008; Ito et al., 2011; Rohr and Wyzga, 2012). Sodium has not been frequently implicated in previous epidemiologic and toxicological studies of $PM_{2.5}$ constituents (Rohr and Wyzga, 2012; Schlesinger, 2007), although one study reported that long-term average sodium ion concentrations partially explained variability in the association between emergency admissions and $PM_{2.5}$ across 26 communities (Zanobetti et al., 2009). Mar et al. (2006) examined sources of pollution and reported associations between sea salt, a sodium-containing source, and mortality. Some time-series studies have reported associations of adverse health outcomes with sulfate (Cao et al., 2012; Ito et al., 2011; Kim et al., 2012; Ostro et al., 2007; Zanobetti et al., 2009), nitrate (Cao et al., 2012; Ito et al., 2011; Kim et al., 2012; Ostro et al., 2007; Peng et al., 2009), and ammonium (Cao et al., 2012; Peng et al., 2009); however, studies have also found sulfate, nitrate, and ammonium to be less toxic than other constituents [e.g., sulfate (Bell et al., 2009; Peng et al., 2009; Tolbert et al., 2007), nitrate (Bell et al., 2009; Darrow et al., 2009; Franklin et al., 2008), ammonium (Bell et al., 2009; Franklin et al., 2008)].

As a sensitivity analysis, we fit a multipollutant model including OCM, EC, silicon, and sodium ion simultaneously and estimated effects that were generally similar in magnitude and direction to single-pollutant model estimates. Previous research has found multipollutant hospitalization effect estimates for EC (Levy et al., 2012) as well as for both EC and OCM (Peng et al., 2009) to be statistically significant. Our multipollutant effect estimates had large standard errors and small posterior probabilities of a positive association, so the possibility of confounding by other constituents has not been completely eliminated. On average across communities, 358 days with

exposure data for all four constituents were included in multipollutant mortality risk models, and some communities had fewer days to estimate multipollutant risks compared to single-pollutant risks, which were estimated from an average of 389 days. In addition, large observed correlations between constituents (e.g., OCM/EC = 0.64) may have affected our model results.

In our analysis of $PM_{2.5}$ total mass and mortality, we found short-term exposure to $PM_{2.5}$ mass was associated with increased mortality, consistent with previous epidemiologic studies (Franklin et al., 2007; Ostro et al., 2006; Zanobetti and Schwartz, 2009). For a 10-μg/m^3 increase in $PM_{2.5}$, we estimated mortality increased 0.38% (95% PI: 0.14, 0.62), whereas other national-level studies found associations of 0.74% (95% CI: 0.41, 1.07) (Franklin et al., 2008), and 0.98% (95% CI: 0.75, 1.22) (Zanobetti and Schwartz, 2009). Although our point estimates were generally smaller than previously reported, methodological differences between our approach and others may explain these differences. To compare estimated $PM_{2.5}$ mass mortality effects with estimated $PM_{2.5}$ constituent effects, we restricted our analysis of $PM_{2.5}$ mass to communities with data from the $PM_{2.5}$ constituent monitoring network, which is a smaller set of communities than studies focusing on $PM_{2.5}$ total mass have previously examined (Dominici et al., 2006; Zanobetti and Schwartz, 2009).

We found little evidence of regional or seasonal variation in associations between mortality and $PM_{2.5}$ constituents or total mass $PM_{2.5}$. Past work has suggested seasonal trends in constituent-specific mortality effects, although results are somewhat ambiguous across studies. Constituent-mortality associations were larger in magnitude during the cooler part of the year than during warmer months in California and in a Chinese city (Huang et al., 2012; Ostro et al., 2007), whereas a study in

New York City reported significant associations of $PM_{2.5}$ constituents with mortality in the warm season but not the cold season (Ito et al., 2011). Silicon and EC were more associated with mortality in the cold season in Seattle, but constituent-mortality associations were similar between seasons in Detroit (Zhou et al., 2011).

In general, the power to detect seasonal and regional differences in $PM_{2.5}$ mass and $PM_{2.5}$ constituent mortality effects in the present study was limited because of the infrequent measurement of the constituent exposures, the relatively short time series, and the small number of ambient monitor locations, particularly in the western United States. Unlike previous studies, we did not find evidence that $PM_{2.5}$ mass mortality effect estimates varied spatially or seasonally (Dominici et al., 2006; Franklin et al., 2007, 2008; Zanobetti and Schwartz, 2009). Model differences may partially explain this discrepancy because earlier seasonal studies used the mean concentration at lags 0 and 1 on season-stratified data (Zanobetti and Schwartz, 2009). In addition, we explicitly tested for seasonal and regional differences using posterior intervals. Peng et al. (2005) documented seasonal and regional variations in estimated effects of PM on mortality, but these estimates were for exposure to PM_{10} ($\leq$ 10 μm in aerodynamic diameter) during an earlier time period (1987-2000). The seasonal and regional differences previously reported may be difficult to observe using more recent data because of declining associations between PM and mortality (Dominici et al., 2007). If seasonal and regional differences in $PM_{2.5}$ mortality effects are explained by differences in the chemical composition of $PM_{2.5}$, we would not expect to find seasonal or regional differences in associations between $PM_{2.5}$ constituents and mortality, which is consistent with our findings. However, in contrast with previous studies, we also did not find evidence of regional or seasonal variation in associations between $PM_{2.5}$ and mortality; consequently, our analysis does

not clarify whether previously observed differences in estimated effects of $PM_{2.5}$ on mortality were driven by differences in chemical composition.

3.4.1 Limitations

We focused on seven constituents that make up the largest fraction of $PM_{2.5}$. However, if $PM_{2.5}$ mass has an effect on mortality that is not mediated through its chemical composition, then we might be more likely to spuriously identify constituents as harmful because they are correlated with $PM_{2.5}$ mass. Future work could apply different regression techniques to distinguish among associations attributable to chemical composition versus $PM_{2.5}$ mass (Mostofsky et al., 2012). In addition, the seven constituents that we evaluated may be correlated with toxic constituents that contribute less to $PM_{2.5}$ by mass. For example, Ito et al. (2004) identified an oil source of PM in New York City that contained nitrate as well as nickel and vanadium, constituents that contribute less to $PM_{2.5}$ by mass, but may be more toxic than more major constituents (Bell et al., 2010; Franklin et al., 2008). However, constituents such as nickel and vanadium often have large proportions of daily data below monitor detection limits (Burnett et al., 2000) and, therefore, may pose additional challenges to analysis. Associations with a given $PM_{2.5}$ chemical component should be considered as potentially indicative of associations with another component or set of components with similar sources.

In our health effects analysis, we did not account for exposure misclassification, which has been demonstrated in previous work (Bell et al., 2011). Depending on the type of measurement error, estimated health effects of estimated community-level exposures may be biased (Zeger et al., 2000). We did not address error resulting from the use of ambient exposure data rather than personal exposure data, which are not

available on the national scale or for long time frames (Dominici et al., 2000). However, a simulation study suggested that improved exposure prediction may not always improve health effect estimation (Szpiro et al., 2011). Using population-weighted community-level exposure data also may not substantially change estimated relative risks (Chang et al., 2011).

Although we performed a sensitivity analysis using different time periods to define seasons, we could not model a smooth transition in the magnitude of associations between pollutants and mortality between consecutive seasons. Further, potential confounders for each season (e.g., weather) may differ by location and may require community-specific modeling approaches. Our approach was to use the same model for each community, and further work may be needed to explore the sensitivity of season-specific estimates to modeling of confounders that vary by location.

Most air pollution health effects studies estimate community-level ambient average pollutant concentrations using the arithmetic mean of monitor concentrations, as we did (Ostro et al., 2007; Peng et al., 2009; Samet et al., 2000b). A previous simulation study suggested that health effect estimates were less biased when the community-level ambient average was estimated using a spatial model rather than the simple arithmetic mean of data from monitors in each community, as we did for the present study (Peng and Bell, 2010). Future work could incorporate spatial modeling to estimate community-level pollutant exposure (Choi et al., 2009). Although distributed lag models are preferred when estimating the effect of pollution over multiple days of exposure (Dominici et al., 2006; Zanobetti and Schwartz, 2009), we could not fit distributed lag models using our non-daily $PM_{2.5}$ constituent data.

3.5 Conclusions

Our analysis substantially builds upon previous studies of PM constituents by providing the first comprehensive national-level assessment of associations between nonaccidental mortality and seven $PM_{2.5}$ constituents in 72 urban communities across the United States during 2000-2005. We found evidence of associations between mortality and OCM, EC, silicon, and sodium ion. We did not find evidence that chemical constituent mortality risks varied by season or region. However, we also did not find evidence of seasonal or regional variation in associations between $PM_{2.5}$ and mortality, in contrast with previous studies. Our study found evidence that some chemical constituents of $PM_{2.5}$ were more associated with mortality than others, which may indicate that regulating PM solely by mass will not sufficiently protect human health.

Table 3.4: National average estimated percent increase in mortality associated with an IQR increase in $PM_{2.5}$ constituents on the same day (lag 0) and two days before (lag 2) for single pollutant models.

	Lag 0		Lag 2	
Pollutant	Estimate (95% PI[a])	P(>0)[b]	Estimate (95% PI[a])	P(>0)[b]
$PM_{2.5}$	0.15 (-0.03, 0.34)	0.95	-0.01 (-0.20, 0.18)	0.44
OCM	-0.04 (-0.38, 0.29)	0.40	0.17 (-0.13, 0.47)	0.87
EC	-0.14 (-0.38, 0.10)	0.13	0.14 (-0.08, 0.36)	0.89
Silicon	0.03 (-0.13, 0.20)	0.65	0.01 (-0.14, 0.17)	0.56
Sodium Ion	-0.01 (-0.17, 0.16)	0.46	0.00 (-0.15, 0.16)	0.51
Nitrate	-0.01 (-0.21, 0.19)	0.46	0.04 (-0.15, 0.24)	0.67
Ammonium	0.11 (-0.20, 0.42)	0.76	-0.06 (-0.38, 0.25)	0.35
Sulfate	0.29 (-0.10, 0.68)	0.93	-0.26 (-0.64, 0.12)	0.09

[a] 95% PI: 95% posterior intervals for the mortality effect estimate.
[b] P(>0): posterior probabilities that the mortality effect estimate is greater than 0

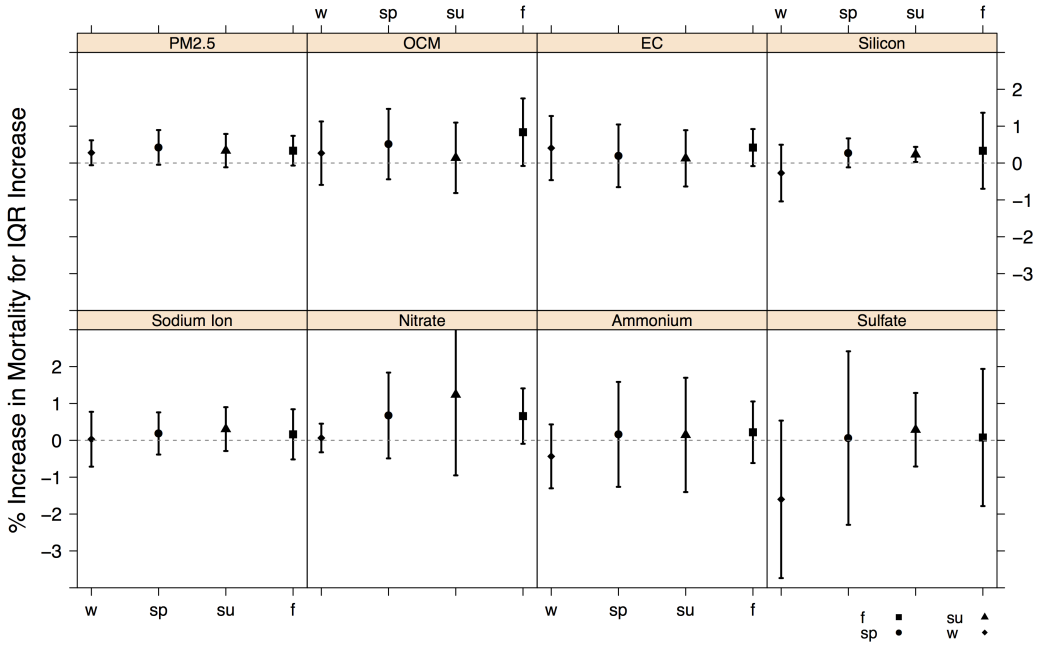

Figure 3.3: Season-specific estimated percent increase in mortality (95% PI) associated with an IQR increase in $PM_{2.5}$ constituents on the previous day for single-pollutant models. Seasons: winter (w: December 21 - March 20), spring (sp: March 21 - June 20), summer (su: June 21 -September 20), fall (f: September 21 - December 20).

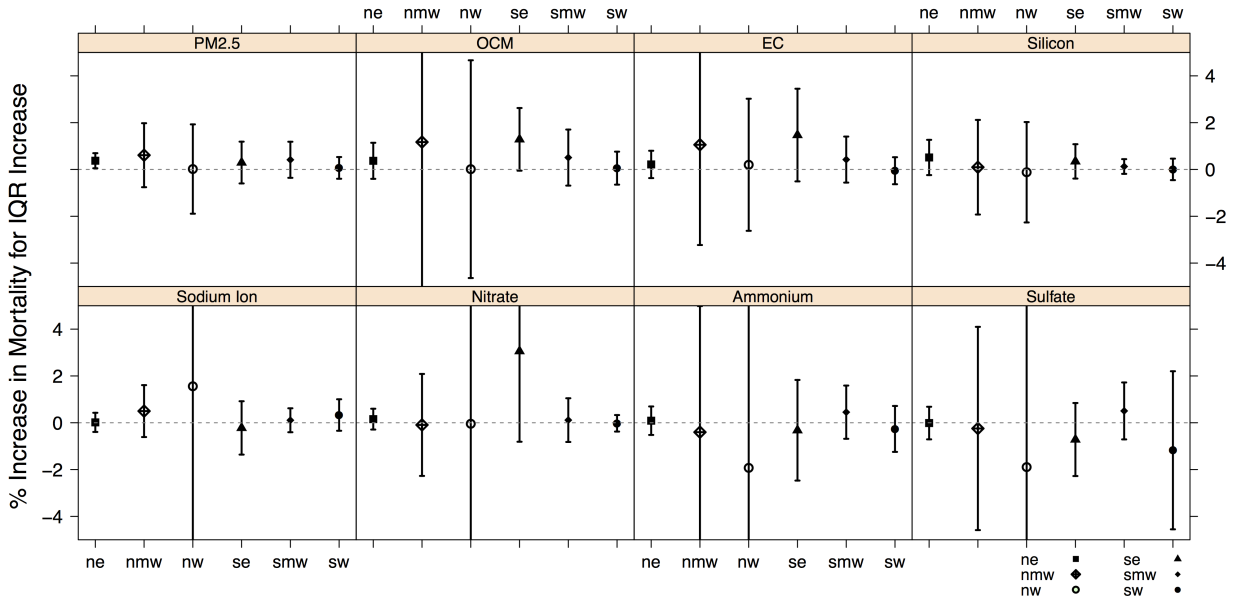

Figure 3.4: Region-specific estimated percent increase in mortality (95% PI) associated with an IQR increase in $PM_{2.5}$ constituents on the previous day for single-pollutant models. Region designations: nw, northeast; nmw, north midwest; nw, northwest; se, southeast; smw south midwest; sw southwest.

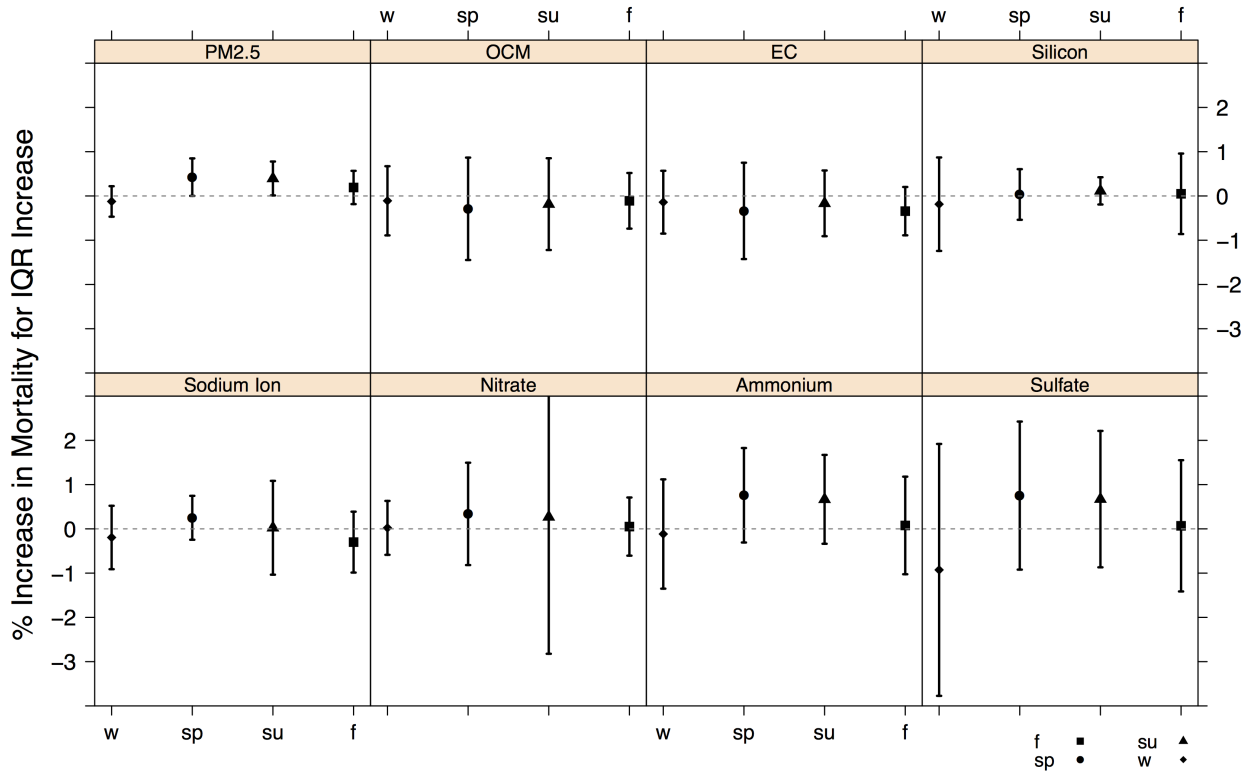

Figure 3.5: Season-specific estimated percent increase in mortality (95% posterior intervals [95% PI]) associated with an IQR increase in $PM_{2.5}$ constituents on the same day (lag 0) for single pollutant models. Seasons are defined: winter (w: December 21 - March 20), spring (sp: March 21 - June 20), summer (su: June 21 -September 20), fall (f: September 21 - December 20).

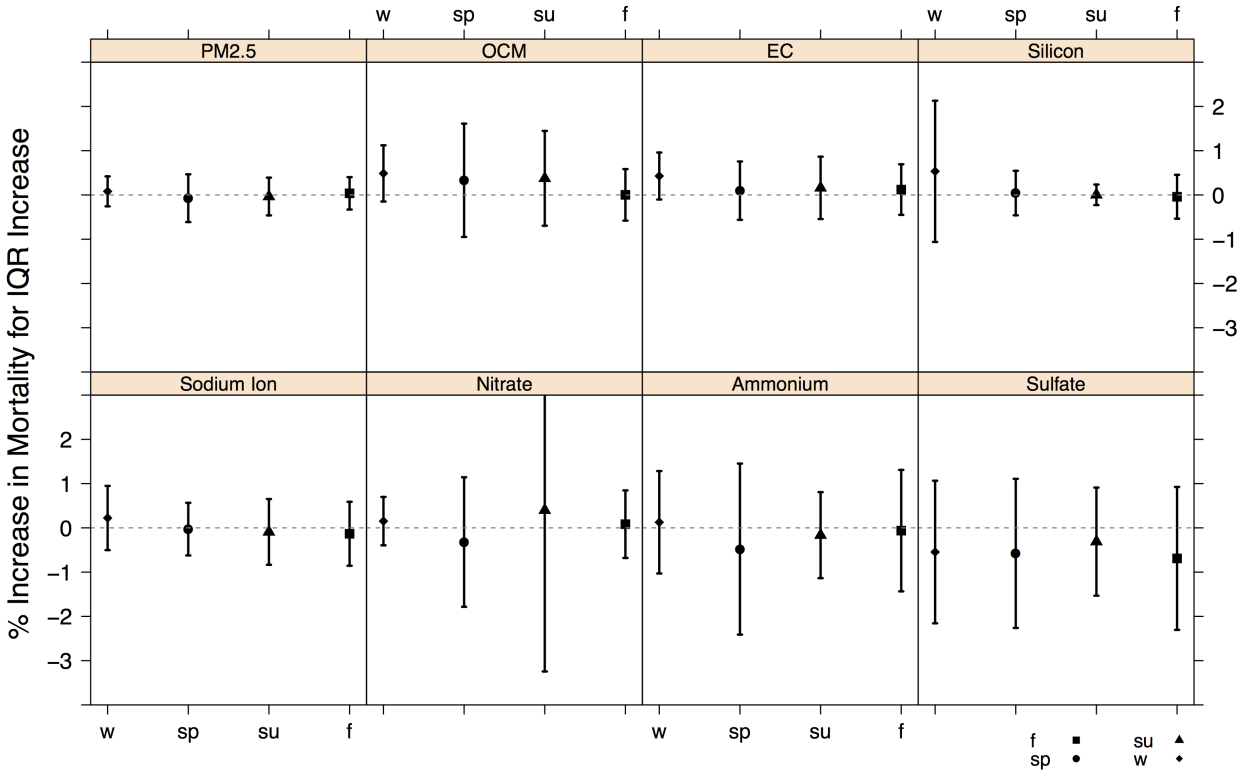

Figure 3.6: Season-specific estimated percent increase in mortality (95% posterior intervals [95% PI]) associated with an IQR increase in $PM_{2.5}$ constituents two days before (lag 2) for single pollutant models. Seasons are defined: winter (w: December 21 - March 20), spring (sp: March 21 - June 20), summer (su: June 21 - September 20), fall (f: September 21 - December 20).

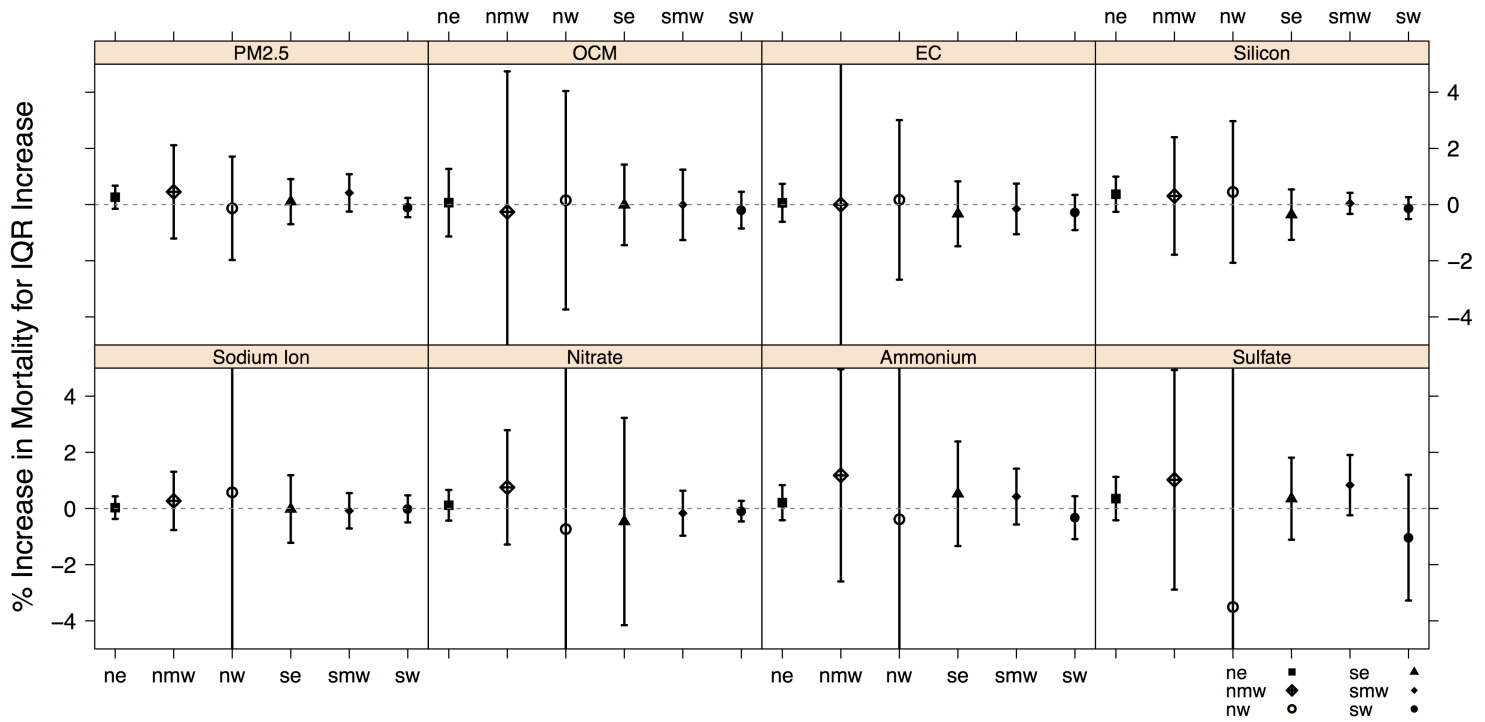

Figure 3.7: Region-specific estimated percent increase in mortality (95% posterior intervals [95% PI]) associated with an IQR increase in $PM_{2.5}$ constituents on the same day (lag 0) for single pollutant models. Region designations include: nw, northeast; nmw, north midwest; nw, northwest; se, southeast; smw south midwest; sw southwest. See Figure 3.1 in the main text for a map of the regions.

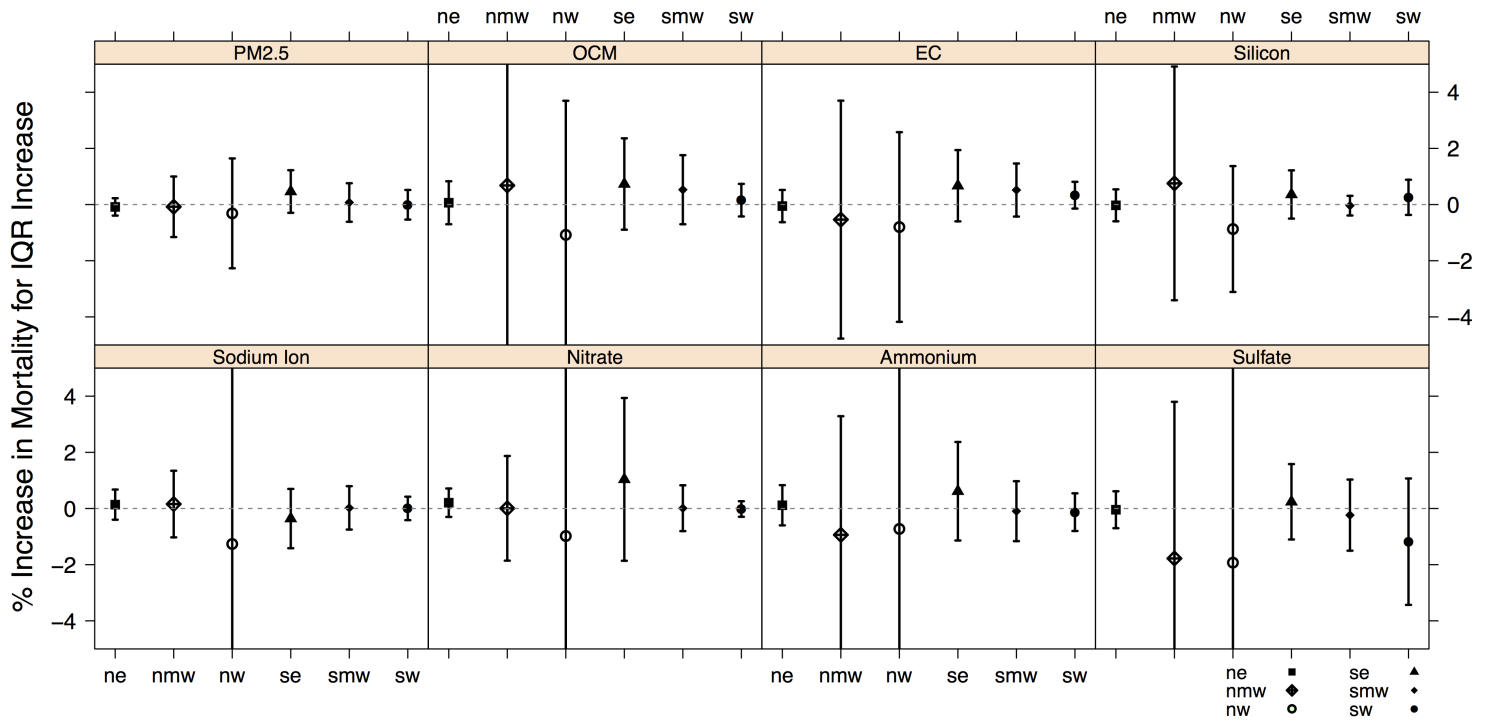

Figure 3.8: Region-specific estimated percent increase in mortality (95% posterior intervals [95% PI]) associated with an IQR increase in $PM_{2.5}$ constituents on two days before (lag 2) for single pollutant models. Region designations include: nw, northeast; nmw, north midwest; nw, northwest; se, southeast; smw south midwest; sw southwest. See Figure 3.1 in the main text for a map of the regions.

Chapter 4

Effects of spatial misalignment for estimating the associations between mortality and particulate matter constituents

Associations between health outcomes and short-term exposure to $PM_{2.5}$ and $PM_{2.5}$ chemical constituents are frequently estimated using time series regression models. Commonly in studies of PM, daily pollutant concentrations are obtained from ambient monitors, whereas health data are aggregated over counties or cities. This difference in spatial resolution between point measures of PM and aggregated health outcomes is referred to as spatial misalignment. To adjust for spatial misalignment, the ambient average PM concentration for a community is traditionally estimated by averaging observed concentrations from monitors within the community. However, the traditional approach may not be a good estimate of the true ambient average for $PM_{2.5}$ chemical constituents, which are spatially heterogeneous and are measured by a sparse network of monitors. Alternatively, spatial models use the spatial variability of the observed data to estimate ambient averages and may lead to different estimated health effects than the traditional approach. In a national-level US study,

we estimated associations between mortality and short-term exposure to $PM_{2.5}$ mass and $PM_{2.5}$ chemical constituents adjusting for spatial misalignment in two ways: using the traditional approach and using a spatial model. The estimated associations between mortality and $PM_{2.5}$ mass and $PM_{2.5}$ constituents from the spatial model were frequently larger in magnitude relative to the traditional approach, indicating that the method used to adjust for spatial misalignment is important in estimating health effects of $PM_{2.5}$ constituents.

4.1 Introduction

In order to estimate health effects of $PM_{2.5}$ in epidemiologic studies, health outcomes are frequently regressed against daily pollutant concentrations. $PM_{2.5}$ concentrations are measured at ambient monitors, however health data (e.g. daily deaths) are frequently aggregated over larger areas such as communities. The difference in spatial resolution between point-level pollution data and aggregated community-level health data is referred to as spatial misalignment. To adjust for spatial misalignment, community-level ambient average $PM_{2.5}$ concentrations are commonly estimated from point-level data. The traditional approach for estimating ambient average $PM_{2.5}$ concentrations in epidemiologic studies is to average observed concentrations from monitors in the community (Samet et al., 2000b; Peng et al., 2009; Ostro et al., 2007; Zhou et al., 2011). When only one $PM_{2.5}$ monitor is available in the community, that monitor is used as a surrogate for the ambient average.

The traditional approach may be a poor measure of the ambient average pollutant concentration when: 1) only a few monitors are available in the community or 2) the pollutant under consideration is spatially heterogeneous (Banerjee et al., 2004; Bell et al., 2007; Peng et al., 2008). These conditions are particularly relevant to the study of health effects associated with $PM_{2.5}$ chemical constituents. The US EPA Chemical Speciation Network (CSN), the national monitoring network measuring $PM_{2.5}$ chemical constituent concentrations, is spatially sparse and many communities have only one speciation monitor (Peng et al., 2009). For communities with speciation monitors, the traditional approach may not yield a good estimate of the true ambient average because monitors are not randomly located throughout communities and

are frequently preferentially placed in areas with high pollution (Environmental Protection Agency, 1999; Özkaynak et al., 2013). For communities with no speciation monitors, the traditional approach cannot be used to estimate ambient average $PM_{2.5}$ constituent concentrations.

Additionally some $PM_{2.5}$ constituents may be more spatially heterogeneous than $PM_{2.5}$ mass because of the local nature of certain sources of $PM_{2.5}$ constituents. $PM_{2.5}$ mass has somewhat large correlations across space and is therefore more spatially homogeneous than other size distributions (Peng et al., 2008), however $PM_{2.5}$ chemical constituents are more spatially heterogenous with lower spatial correlations (Peng and Bell, 2010). Mobile sources of $PM_{2.5}$ such as motor vehicle traffic may drive spatial heterogeneity in associated $PM_{2.5}$ constituents such as EC and OCM. A US study of $PM_{2.5}$ constituents from 2000-2006 found sodium ion and elemental carbon (EC) to be very spatially heterogeneous, even within small distances (Peng and Bell, 2010). The traditional approach for estimating the ambient average implicitly assumes spatial homogeneity of pollutants, which is an incorrect assumption for many $PM_{2.5}$ constituents (Peng and Bell, 2010). While previous studies relating $PM_{2.5}$ total mass to health outcomes have used the traditional approach to adjust for spatial misalignment, studies of the health effects of $PM_{2.5}$ constituents may need a different method to estimate the ambient average.

Spatial models can be used as an alternative to the traditional approach to adjust for spatial misalignment. To estimate the ambient average, spatial models first estimate the spatial correlation of the pollutant concentrations. Once the spatial model has been fitted to the available data, it can be used to estimate the daily community-level ambient average concentration for the pollutant. While the traditional approach

only uses monitors inside the community to estimate the ambient average concentration, spatial models can use all available monitoring data. Furthermore, predictions from spatial models are more robust to outlying monitor concentrations. The estimated community-level ambient average from the spatial model is less dependent on individual monitor values, resulting in less temporal variability than observed monitor concentrations. The reduction in temporal variability subsequently reduces the amount of statistical information available in the health effects regression model and better reflects the uncertainty in estimating ambient average concentrations.

The spatial correlation of $PM_{2.5}$ chemical constituents may vary across the US since some constituents are generated by spatially varying sources. Previous analyses of $PM_{2.5}$ speciation data have shown that there is significant spatial variation in constituent concentrations across the US (Bell et al., 2007), which may indicate presence of different sources or relative differences in the source contributions. $PM_{2.5}$ constituents may have different spatial correlation structures depending on the sources present in the region. For example, $PM_{2.5}$ zinc could be generated by an incineration source (Ito et al., 2004), a metals-related industrial source (Thurston et al., 2011), or motor vehicle traffic (Bell et al., 2010). Since the transport of zinc may depend on the generating source, the spatial correlation of zinc may depend on what proportion is attributable to different sources. One previous study used a stationary spatial model to estimate ambient average $PM_{2.5}$ constituent concentrations in the US (Peng and Bell, 2010). However, a nonstationary spatial model would allow the spatial correlation to vary between regions of the US with different sources of $PM_{2.5}$. Because the available data from the EPA CSN are spatially and temporally sparse, there is not enough data to fit previously proposed nonstationary models (Fuentes and Smith,

2001; Higdon, 1998; Paciorek and Schervish, 2006). We incorporated a nonstationary element into our spatial modeling approach by fitting separate spatial models to regions in the US that share similar $PM_{2.5}$ sources.

We estimated associations between mortality and short-term exposure to total mass $PM_{2.5}$ and seven major $PM_{2.5}$ chemical constituents using two approaches to adjust for spatial misalignment: a traditional approach and a spatial model. We fitted spatial models separately for six regions in the US for $PM_{2.5}$ total mass and each $PM_{2.5}$ constituent. If the models fitted to each region differ substantially for a pollutant, the results would provide evidence that fitting a stationary spatial model across the US may not be adequate and a nonstationary spatial model may better represent the spatial correlation of the pollutant. We compared estimated ambient average pollutant concentrations between the traditional approach and our spatial model for 72 US communities. Using both methods to adjust for spatial misalignment, we estimated associations between all-cause mortality and total mass $PM_{2.5}$ and $PM_{2.5}$ constituents for 72 communities. In this work, we compared estimated ambient average concentrations and estimated associations with mortality between a traditional approach and a spatial model to adjust for spatial misalignment.

4.2 Data

For the period 2000-2005, we obtained daily $PM_{2.5}$ chemical constituent concentrations from the US EPA CSN (Bell et al., 2007), which is a national monitoring network of approximately 250 ambient speciation monitors located throughout the US. While the monitors measure concentrations for $PM_{2.5}$ and over 50 chemical constituents of $PM_{2.5}$, we restricted our analysis to seven constituents that make up the

largest fraction of $PM_{2.5}$ by mass or are highly correlated with $PM_{2.5}$ total mass: sulfate, nitrate, sodium ion, silicon, ammonium, organic carbon matter (OCM) and elemental carbon (EC). On average, these seven constituents together form 79-85% of total $PM_{2.5}$ by mass, while each of the remaining constituents of $PM_{2.5}$ contribute less than 1% to the total mass (Bell et al., 2007). The ambient monitors in the EPA CSN generally measure concentrations every one in six days. We also obtained daily concentrations of $PM_{2.5}$ total mass from the larger US EPA Air Quality System (AQS), which includes approximately 1,400 monitoring sites (Peng et al., 2009; Dominici et al., 2006). The AQS monitoring network has been previously applied in studies of the health effects of total mass $PM_{2.5}$ (Peng et al., 2009; Zanobetti and Schwartz, 2009).

We obtained daily all-cause mortality (excluding accidental deaths) aggregated from US death certificate data from the National Center for Health Statistics. We limited the mortality data to daily deaths in 2000-2005 for 72 communities, where each community is a county or set of counties including an urban area. As in a previous analysis (Krall et al., 2013), each of these 72 communities are located in the continental US, have a $PM_{2.5}$ chemical speciation monitor within its boundaries, and had sufficient data for fitting time series regression models (Chapter 3). The communities are listed in Table 4.1 and are shown in Figure 4.1. We also obtained daily temperature for each community from the National Oceanic and Atmospheric Association.

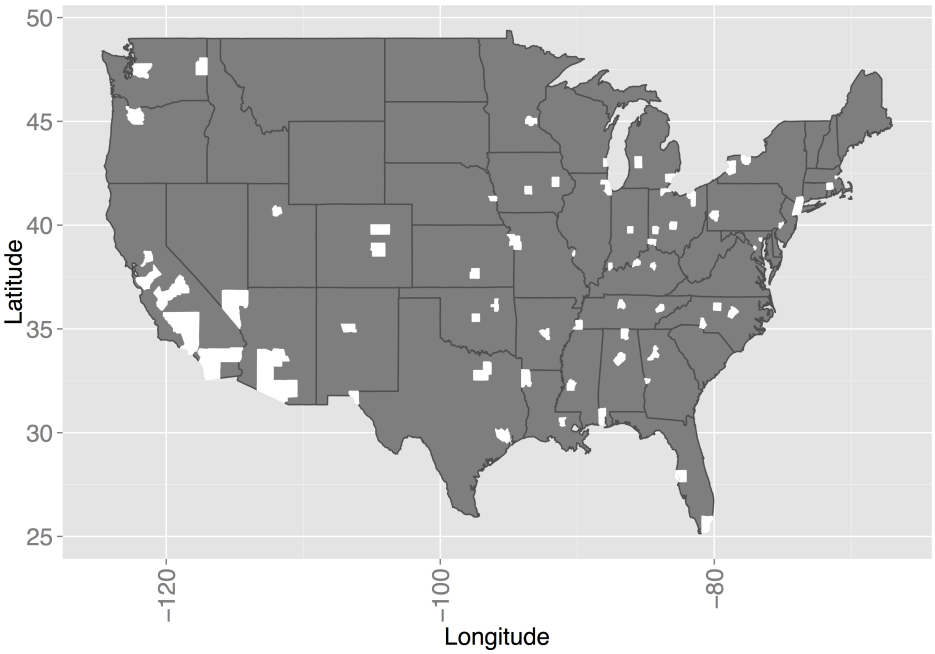

Figure 4.1: Map of the continental US showing locations of the 72 urban communities used in this analysis.

Table 4.1: Communities used in this analysis

Akron, OH	Albuquerque, NM	Atlanta, GA
Bakersfield, CA	Baltimore, MD	Baton Rouge, LA
Birmingham, AL	Boston, MA	Buffalo, NY
Cedar Rapids, IA	Charlotte, NC	Chicago, IL
Cincinnati, OH	Cleveland, OH	Colorado Springs, CO
Columbus, GA	Columbus, OH	Dallas/Fort Worth, TX
Dayton, OH	Denver, CO	Des Moines, IA
Detroit, MI	El Paso, TX	Evansville, IN
Fresno, CA	Grand Rapids, MI	Greensboro, NC
Houston, TX	Huntsville, AL	Indianapolis, IN
Jackson, MS	Kansas City, KS	Kansas City, MO
Knoxville, TN	Las Vegas, NV	Lexington, KY
Little Rock, AR	Los Angeles, CA	Louisville, KY
Memphis, TN	Miami, FL	Milwaukee, WI
Minneapolis/St. Paul, MN	Mobile, AL	Modesto, CA
Nashville, TN	New York, NY	Oklahoma City, OK
Omaha, NE	Philadelphia, PA	Phoenix, AZ
Pittsburgh, PA	Portland, OR	Providence, RI
Raleigh, NC	Riverside, CA	Rochester, NY
Sacramento, CA	Salt Lake City, UT	San Diego, CA
San Jose, CA	Seattle, WA	Shreveport, LA
Spokane, WA	St. Louis, MO	St. Petersburg, FL
Tampa, FL	Toledo, OH	Tucson, AZ
Tulsa, OK	Washington, DC	Wichita, KS

4.3 Methods

4.3.1 Estimating ambient pollutant concentrations

To adjust for spatial misalignment between pollution and health data, we estimated ambient average concentrations of seven $PM_{2.5}$ constituents and $PM_{2.5}$ total mass using two approaches. First, we used the traditional approach frequently applied in time-series studies of the health effects of $PM_{2.5}$ and $PM_{2.5}$ constituents (Krall et al., 2013; Peng et al., 2009). The traditional approach estimates daily ambient average

pollutant concentrations for each community by averaging daily concentrations observed at ambient monitors inside the community. For communities with only one monitor, the traditional approach uses daily concentrations from that monitor. We also estimated ambient average pollutant concentrations using a spatial model, which is described in detail in the following sections.

Spatial model

For the spatial model, we first divided our ambient monitoring data into several geographic regions representing similar sources of $PM_{2.5}$. We modeled the concentrations of each of the seven $PM_{2.5}$ constituents and $PM_{2.5}$ total mass as separate Gaussian processes, $x_{r(s)}(s,t)$, observed at location s and day t, where $r(s)$ is the region corresponding to location s. For each pollutant, the Gaussian process consists of a region-specific mean, $\mu_{r(s)}$, and two additional region-specific terms, $w_{r(s)}(s,t)$ and $\varepsilon_{r(s)}(s,t)$:

$$x_{r(s)}(s,t) = \mu_{r(s)} + w_{r(s)}(s,t) + \varepsilon_{r(s)}(s,t)$$

The second term $\varepsilon_{r(s)}(s,t)$ is a mean zero, white noise process with variance $\tau^2_{r(s)}$. Our ability to estimate small distances is constrained by the distance between the two closest monitors, d_{min}. Variations in air pollution over distances smaller than d_{min} (microscale variation) cannot be captured by observed monitor concentrations. The variance of $\varepsilon_{r(s)}(s,t)$ accounts for microscale variation or measurement error (Paciorek and Schervish, 2006).

The Gaussian process $w_{r(s)}(s,t)$ is also mean zero and accounts for spatial variation with covariance

$$Cov\left\{w_{r(s)}(s,t), w_{r(s')}(s',t')\right\} = \begin{cases} \sigma^2_{r(s)}\rho\left(||s-s'||;\phi_{r(s)},\kappa_{r(s)}\right) & t = t' \\ 0 & t \neq t' \end{cases}$$

where $||s-s'||$ is the distance between two spatial locations and $\sigma^2_{r(s)}$ is the variance of $w_{r(s)}(s,t)$. For two spatial locations, s and s', the covariance term $\rho\left(||s-s'||;\phi_{r(s)},\kappa_{r(s)}\right)$ is a Matérn covariance function with Bessel function of the third kind, $\mathscr{K}$:

$$\rho\left(||s-s'||;\phi_{r(s)},\kappa_{r(s)}\right)=\begin{cases}\frac{1}{2^{\kappa_{r(s)}-1}\Gamma(\kappa_{r(s)})}\left(\frac{||s-s'||}{\phi_{r(s)}}\right)^{\kappa_{r(s)}}\mathscr{K}_{\kappa_{r(s)}}\left(\frac{||s-s'||}{\phi_{r(s)}}\right) & s\neq s'\\ 1 & s=s'\end{cases} \tag{4.1}$$

We used the Matérn covariance function because it is flexible and allows for an estimate of $\kappa_{r(s)}$, which represents the smoothness of the model. The remaining term $\phi_{r(s)}$ is the range parameter for the Matérn function and describes the decay of the spatial correlation with increasing distance.

Within each region, our spatial model is stationary because it depends only on the distance $||s-s'||$. However, each parameter in our model depends on $r(s)$, which allows the covariance function to differ by region and accounts for potential nonstationarity. A nonstationary spatial model may better represent $PM_{2.5}$ constituents because the spatial correlation of $PM_{2.5}$ constituents likely varies across the US with varying meteorological conditions and varying sources of $PM_{2.5}$. Additionally, this proposed nonstationary spatial model can be fit to the available $PM_{2.5}$ constituent data, which are sparse temporally and spatially.

Under the Gaussian process, we modeled the observed data for each pollutant in region $r(s)$, designated by $\mathbf{m}_{r(s)}$, as jointly normal. To estimate parameters $\mathbf{\Theta}_{r(s)} = (\sigma_{r(s)},\ \phi_{r(s)},\ \kappa_{r(s)},\ \mu_{r(s)},\ \tau_{r(s)})$ for each pollutant in each region $r(s)$, we maximized the corresponding likelihood:

$$L\left(\mathbf{\Theta}_{r(s)};\mathbf{m}_{r(s)}\right)\propto\prod_{t=1}^{T}|\boldsymbol{M}_{r(s)}|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{m}_{\mathbf{r(s)}}(t)-\mu_{r(s)}\mathbf{1})'\boldsymbol{M}^{-1}_{r(s)}(\mathbf{m}_{r(s)}(t)-\mu_{r(s)}\mathbf{1})\right\} \tag{4.2}$$

where $\mathbf{M}_{r(s)}$ is the appropriate covariance matrix with (i,k) element equal to

$$\sigma^2_{r(s)}\rho(||s_i - s_k||; \phi_{r(s)}, \kappa_{r(s)}) + \tau^2_{r(s)} \cdot I_{s_i=s_k} \tag{4.3}$$

and s_i and s_k are monitor locations in region $r(s)$. Additionally $\mu_{r(s)}\mathbf{1}$ is a vector with the same length as $\mathbf{m}_{r(s)}(t)$. To maximize the normal likelihood based on the spatial model and to estimate the model parameters for each region and pollutant, we used standard nonlinear optimization techniques. Specifically, we used the limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method for optimization, which allows specification of lower and upper bounds for the parameters (Byrd et al., 1995).

Estimation of ambient concentrations using the spatial model

For each $PM_{2.5}$ constituent and total mass $PM_{2.5}$, we used the fitted spatial models to estimate the ambient average pollutant concentration $X(c,t)$ for day t in each community c. Using the conditional normal induced by the Gaussian process, we estimated the average pollutant concentration conditional on all reporting monitors in the corresponding region $r(c)$. For each pollutant, the predictive distribution is

$$X(c,t) \mid \mathbf{m}_{r(c)}(t) \sim N\left(\mu_c + \mathbf{h}'_{r(c)}\mathbf{M}^{-1}_{r(c)}(\mathbf{m}_{r(c)}(t) - \mu_{r(c)}\mathbf{1}),\ \psi_{r(c)} - \mathbf{h}'_{r(c)}\mathbf{M}^{-1}_{r(c)}\mathbf{h}_{r(c)}\right) \tag{4.4}$$

where $\mu_c = \mu_{r(c)}$ is the pooled community-wide ambient average. From the likelihood in equation 4.2, $\mathbf{m}_{r(c)}(t)$ is normally distributed with mean $\mu_{r(c)}\mathbf{1}$ and covariance $\mathbf{M}_{r(c)}$. The term $\mathbf{M}_{r(c)}$ has elements (i,k) as in equation 4.3 with distances $||s_i - s_k||$ representing the distances between all reporting monitors in the region $r(c)$. The variance $\psi_{r(c)}$ is a Monte Carlo integral, computed by summing equation 4.3 over a dense grid of points throughout the community.

The remaining variable $\mathbf{h}_{r(c)}$ is the vector of covariances between locations in the community and each reporting monitor l in region $r(c)$. The l^{th} element of $\mathbf{h}_{r(c)}$ is

$$\sigma^2_{r(c)} \int_{A(c)} \rho\left(||m_{r(c),l}(t) - s||; \phi_{r(c)}, \kappa_{r(c)}\right) ds + \tau^2_{r(c)}$$

where $A(c)$ is the area spanned by community c and $m_{r(c),l}(t)$ is the observed pollutant concentration at monitor l. Because this integral is difficult to evaluate analytically, we approximated the integral using Monte Carlo integration.

We used the conditional mean in equation 4.4 to estimate daily ambient average concentrations for community c and each pollutant. For a given spatial location, observations from monitors far from that location influence the predicted concentration less than closer observations. If the estimated Gaussian process is not smooth and has a small range, the information that can be gained from surrounding monitors is minimal. In this case, the elements of $\mathbf{h}_{r(c)}$ would be close to 0 and the community-level pollutant concentration is estimated by the estimated pollutant mean in the region, μ_c. Since our ability to estimate the temporal correlation is limited by the sampling scheme of the ambient monitors, the spatial model assumes a priori that pollution concentrations are independent with respect to time, conditional on the mean. However, temporal correlation in the raw data may result in temporal correlation in predictions generated by this model. Therefore, the predictions used in the mortality analysis will not necessarily be temporally independent.

4.3.2 Mortality Analysis

We estimated associations between mortality and short-term exposure to $PM_{2.5}$ and $PM_{2.5}$ chemical constituents with overdispersed Poisson regression models. As independent variables in our mortality risk models, we used estimated ambient average

pollutant concentrations from both the traditional approach and the spatial model. For each community, we modeled the association between each pollutant j and mortality with a single pollutant model,

$$\log\left\{E\big(\,Y(t)\mid X_j(t),\mathbf{Z}(t)\big)\right\}=\beta_0+X_j(t)\beta_j+\mathbf{Z}(t)'\boldsymbol{\alpha} \tag{4.5}$$

where for each day t, $X_j(t)$ is the predicted pollutant concentration for pollutant j, $Y(t)$ is the number of deaths, and $\mathbf{Z}(t)$ is a vector of potential confounders (e.g. temperature). For each pollutant j, β_j is the log relative risk of mortality for a 1 $\mu g/m^3$ increase in pollutant concentration.

To estimate national-level associations between mortality and each pollutant, we combined the community-specific estimated log relative risks using a two-level normal Bayesian hierarchical model (Everson and Morris, 2000) as in previous work (Krall et al., 2013; Peng et al., 2009). We report results as the percent increase in mortality for an interquartile range (IQR) increase in the pollutant concentration.

We compared community-specific mortality risk estimates between ambient concentrations estimated using the spatial model and the traditional approach. We regressed the spatial model mortality risk estimates on the estimates from the traditional approach, which we weighted by the inverse variances of the spatial model estimates,

$$\bar{\beta}_{c,j}=\gamma_j+\nu_j\,\hat{\beta}_{c,j}+\delta_{c,j} \tag{4.6}$$

In equation 4.6 for pollutant j and community c, $\bar{\beta}_{c,j}$ is the community-specific mortality risk estimate using the spatial model and $\hat{\beta}_{c,j}$ is the weighted community-specific mortality risk estimate using the traditional approach. We are primarily interested in ν_j, which describes the relationship between mortality risk estimates from the spatial model and the traditional approach.

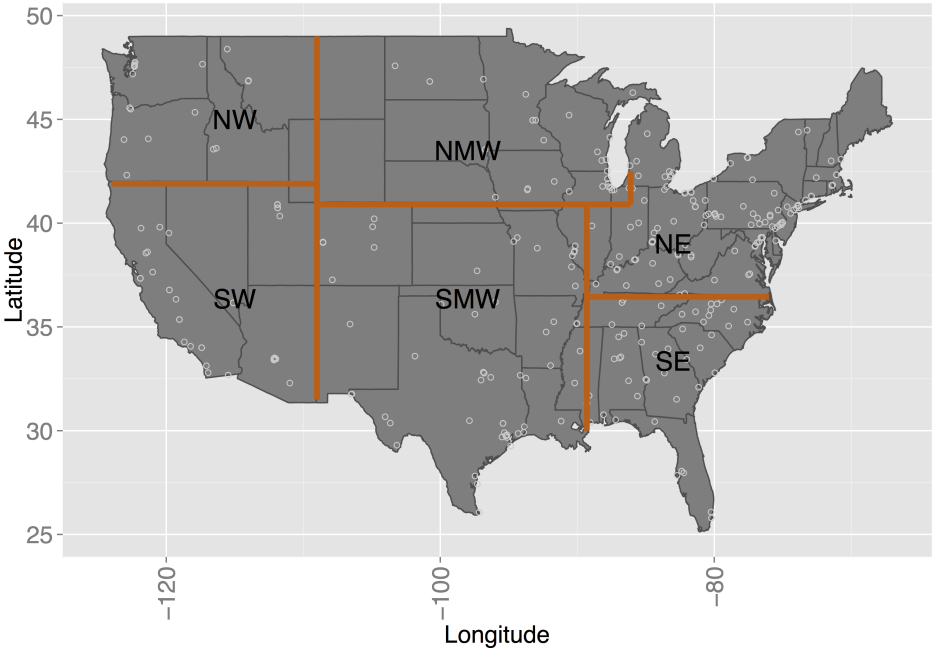

Figure 4.2: Map of the continental US showing locations of $PM_{2.5}$ speciation monitors in the US EPA CSN and regions used to fit the spatial models.

4.4 Results

4.4.1 Spatial models

We fitted spatial models separately for six regions in the US, which were loosely based on regions used by the US EPA (Environmental Protection Agency, 2009) and on the regional distribution of $PM_{2.5}$ sources (Peng et al., 2005; Samet et al., 2000b; Zanobetti and Schwartz, 2009). The regions (and respective number of $PM_{2.5}$ constituent monitors) consist of the northeast (NE, 119), southeast (SE, 48), north midwest (NMW, 31), south midwest (SMW, 58), northwest (NW, 22), and southwest (SW, 35). Figure 4.1 shows the locations of the $PM_{2.5}$ speciation monitors and the regions used to fit the spatial models.

To decrease the impact of outlying observed concentrations on the spatial model

fit, we only used days with at least 20 observations from monitors in the region. Because the northwest region only had 22 speciation monitors, we used all days with at least 5 observations in the northwest. We log-transformed daily concentrations to make the data for each pollutant more symmetric and eliminated days with zero concentrations (less than 3% of daily concentrations). For each pollutant at each monitor, we detrended the data to remove seasonal and long-term trends by fitting a linear model with a categorical variable for day of week and a smooth function of time with 7 degrees of freedom (df) per year. We excluded monitors with fewer than 50 observations to adequately detrend the data, which reduced the number of monitors to fit the spatial model. To create one concentration time series for each spatial location, we averaged daily pollutant concentrations for collocated monitors. Distance between monitors was measured in kilometers by Meeus distance, which accounts for the elliptical shape of the earth. The median distance between monitors ranged from 205 km to 763 km across regions and the largest distance was in the south midwest (1667 km).

We fitted spatial models for each pollutant and region described in equation 4.1 and estimated the parameters $\boldsymbol{\Theta}_{r(s)}=(\sigma_{r(s)},\ \phi_{r(s)},\ \kappa_{r(s)},\ \mu_{r(s)},\ \tau_{r(s)})$ using L-BFGS-B optimization. We chose bounds to ensure that all parameters except $\mu_{r(s)}$ were positive. Figure 4.3 shows the fitted Matérn correlation functions for each pollutant across the six regions. Most pollutants had spatial correlations less than 0.5 at distances over 500 km. At distances less than 100 km, OCM in the northwest had very little correlation while silicon had high correlations across all regions. Sulfate and nitrate, which are generated primarily by regional $PM_{2.5}$ sources, show less nonstationarity than other constituents such as OCM, which is a mobile source generated primarily by motor vehicle traffic.

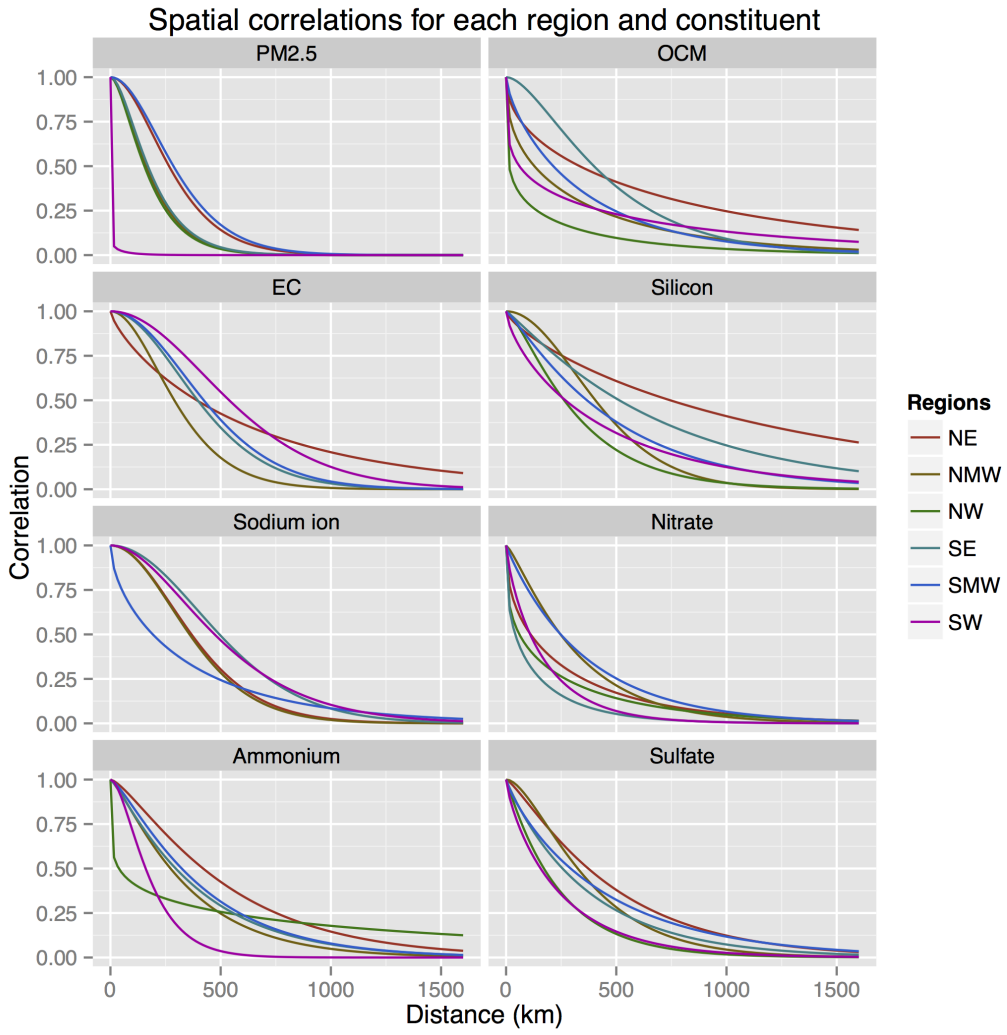

Figure 4.3: Region-specific estimated Matérn correlations for each pollutant for distances measured in kilometers.

4.4.2 Estimation of ambient averages

We estimated the daily ambient average pollutant concentrations using the traditional approach and the spatial model. For the spatial model, we predicted concentrations at approximately 5000 randomly chosen points throughout each community using the mean of the conditional normal in equation 4.4. To obtain estimates of the ambient average, we averaged these spatial model predictions across the 5000 points. We estimated ambient average concentrations using the spatial model for days with at least 3 observed pollutant concentrations in the region. Because fewer data were needed to make predictions than to fit the spatial model, we did not use the same restriction as fitting the spatial model (all days with $>$20 observations in the NE, SE, NMW, SMW, SE; all days with $>$5 observations in the NW). We estimated ambient average concentrations for the same set of days for both the traditional approach and the spatial model to eliminate the possibility that differences between the methods were attributable to differences in the days of observation. For each community, we estimated ambient averages on days with 1) at least one observation in the community so the traditional approach could be applied and 2) at least 3 observations in the region so the spatial prediction could be made.

Table 4.2 gives the mean of monitor-specific IQRs, the number of days of estimated ambient average concentrations for $PM_{2.5}$ total mass and $PM_{2.5}$ constituents, and the average ambient concentration across communities for both the spatial model and the traditional approach. On average across communities, the spatial model estimated smaller average ambient concentrations than the traditional approach (Table 4.2). Community-level ambient averages from the traditional approach and the spatial model differed most in large communities with few monitors and for spatially

heterogeneous pollutants. For both the traditional approach and the spatial model, Figure 4.4A shows time series plots for nitrate in Pittsburgh, a small community (730 square miles) with three speciation monitors. The difference between the two time series for nitrate, a relatively spatially homogeneous constituent, is small and we would expect estimated mortality effects to be similar between the traditional approach and the spatial model. Figure 4.4B has corresponding time series plots for EC in Los Angeles, a large community (4,058 square miles) with only one monitor. The time series is substantially more variable for the traditional approach, which does not account for the high spatial variability of EC and the low monitor coverage.

Table 4.2: Median of monitor-specific IQRs, mean (minimum, maximum) number of days with estimated ambient concentrations, and mean (minimum, maximum) ambient average concentrations for pollutants estimated using the spatial model and the traditional approach.

Pollutant	IQR	Days		Spatial		Traditional	
$PM_{2.5}$	8.00	1631	(456, 2190)	12.4	(6.39, 16.43)	13.48	(6.38, 22.84)
OCM	3.08	376	(58, 760)	3.75	(2.43, 6.48)	4.33	(2.30, 8.68)
EC	0.37	382	(58, 767)	0.56	(0.31, 0.84)	0.73	(0.30, 1.52)
Silicon	0.08	382	(56, 775)	0.10	(0.05, 0.29)	0.13	(0.05, 0.51)
Sodium ion	0.11	360	(58, 689)	0.11	(0.04, 0.29)	0.15	(0.04, 0.49)
Nitrate	1.22	375	(58, 668)	1.50	(0.54, 4.48)	1.91	(0.51, 9.51)
Ammonium	1.14	379	(58, 779)	1.24	(0.30, 1.97)	1.48	(0.32, 3.72)
Sulfate	2.75	379	(58, 779)	2.99	(0.65, 5.27)	3.26	(0.66, 5.74)

For each community and pollutant, we computed the root mean squared difference (RMSD) between the ambient average estimated using the spatial model, $\bar{x}_s$, and the ambient average from the traditional approach, $\bar{x}_t$, $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\bar{x}_{s,i}-\bar{x}_{t,i})^2}$, where n is the number of days of data. We scaled the RMSD by the average pollutant concentration from the raw data so that we could compare the spatial distribution of measurement error between pollutants. In Figure 4.5, we overlaid the RMSD results for each pollutant on a map of the US to illustrate which communities have

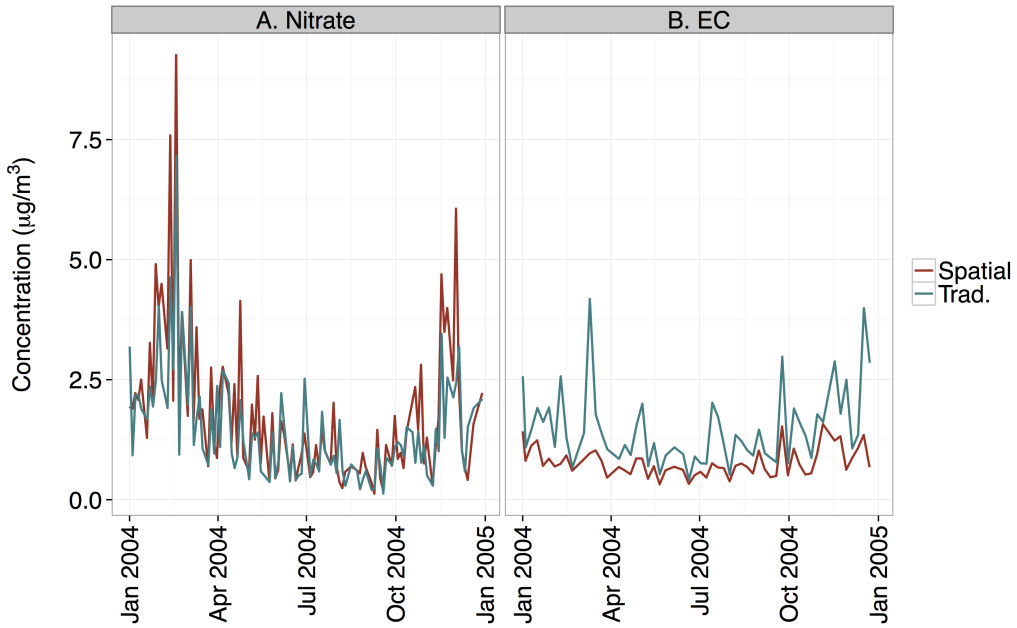

Figure 4.4: Time series plots of estimated ambient concentrations using the traditional approach and the spatial model for A. nitrate in Pittsburgh, PA and B. EC in Los Angeles, CA.

the greatest differences in estimated ambient averages between the spatial model and the traditional approach. $PM_{2.5}$ and sulfate have the smallest RMSD across communities, while both sodium ion and silicon have communities in the central US with large RMSDs. Nitrate and ammonium have larger RMSDs in the northeast and in the southwest, while sulfate has larger RMSDs in the northeast.

4.4.3 Mortality Analysis

We modeled the associations between mortality and short-term exposure to $PM_{2.5}$ total mass and $PM_{2.5}$ constituents using overdispersed Poisson regression models. As potential confounders ($\mathbf{Z}(t)$ in equation 4.5), we included smooth functions (natural spline) of temperature (df=3), one-day lag of temperature (df=3), time (8 df per year), as well as categorical variables for age (under 65, 65-74, and 75 years and older) and day of week. We estimated associations with mortality for exposure to $PM_{2.5}$ constituents and total mass $PM_{2.5}$ on the same day (lag 0), previous day (lag 1) and two days before (lag 2). Since the EPA CSN network measures constituent concentrations every 1-in-6 days we were unable to fit distributed lag models.

Estimated mortality effects and 95% posterior intervals (95% PI) for lags 0, 1, and 2 using ambient concentrations estimated from both the spatial model and the traditional approach are shown in Figures 4.6, 4.7, and 4.8. All results are reported as the percent increase in mortality for an IQR increase in the pollutant, where IQRs were computed as the median of monitor-specific IQRs (Table 4.2). As in previous work (Zanobetti and Schwartz, 2009), we found evidence of positive associations between previous-day exposure to total mass $PM_{2.5}$ and mortality. For an IQR increase in previous day total mass $PM_{2.5}$, we estimated an increase in mortality of 0.29% (95% PI: 0.10%, 0.49%) using the traditional approach. For the spatial model, we

Figure 4.5: Root mean squared differences between the estimated ambient concentrations from the spatial model and and the traditional approach, scaled by mean of the pollutant from the raw data.

estimated an increase in mortality of 0.44% (95% PI: 0.22%, 0.67%) for an IQR increase in previous day $PM_{2.5}$.

For all $PM_{2.5}$ constituents, the association between previous-day exposure and mortality was greater for the spatial model than for the traditional approach. At lag 1, we found strongest evidence that mortality was associated with increases in OCM, EC, and silicon. An IQR increase in previous-day OCM was associated with increases in mortality of 0.37% (95% PI: 0.05%, 0.69%) for the traditional approach and 0.60% (95% PI: 0.10%, 1.11%) for the spatial model. Similarly for the traditional approach and spatial model, an IQR increase in EC at lag 1 was associated with mortality increases of 0.21% (95% PI: -0.02%, 0.43%) and 0.77% (95% PI: 0.22%, 1.31%) respectively. Silicon was significantly associated with mortality at lag 1 for the traditional approach (0.16%, 95% PI: 0.02%, 0.30%) and the association was larger in magnitude, but not statistically significant, for the spatial model (0.22%, 95% PI: -0.08%, 0.52%). In addition, we found some evidence of associations between lag 2 exposure to OCM and EC and mortality for the spatial model, though associations were smaller in magnitude than for previous day exposure. We did not find evidence of associations at any lag for sodium ion, nitrate, ammonium, or sulfate. Estimated mortality effects using the spatial model were generally greater in magnitude than those using the traditional approach for estimating ambient averages, although estimates from the spatial model had larger standard errors.

We compared the community-specific mortality risk estimates between ambient averages from the spatial model and the traditional approach using the simple linear regression in equation 4.6. The linear regression coefficients ν_j from equation 4.6 for each pollutant j and lag are displayed in table 4.3. The majority of ν_j's were greater than one, ranging from 0.9 for nitrate at lag 2 to 1.42 for silicon at lag 0. Intercepts

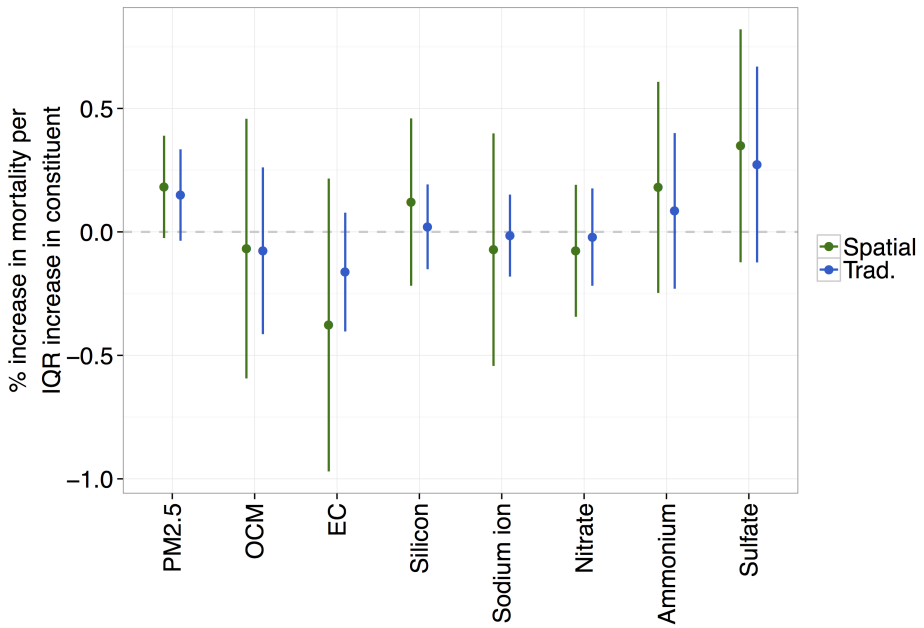

Figure 4.6: Estimated percent increase in mortality (95% PI) associated with an IQR increase in $PM_{2.5}$ mass and $PM_{2.5}$ constituents for same-day exposure (lag 0).

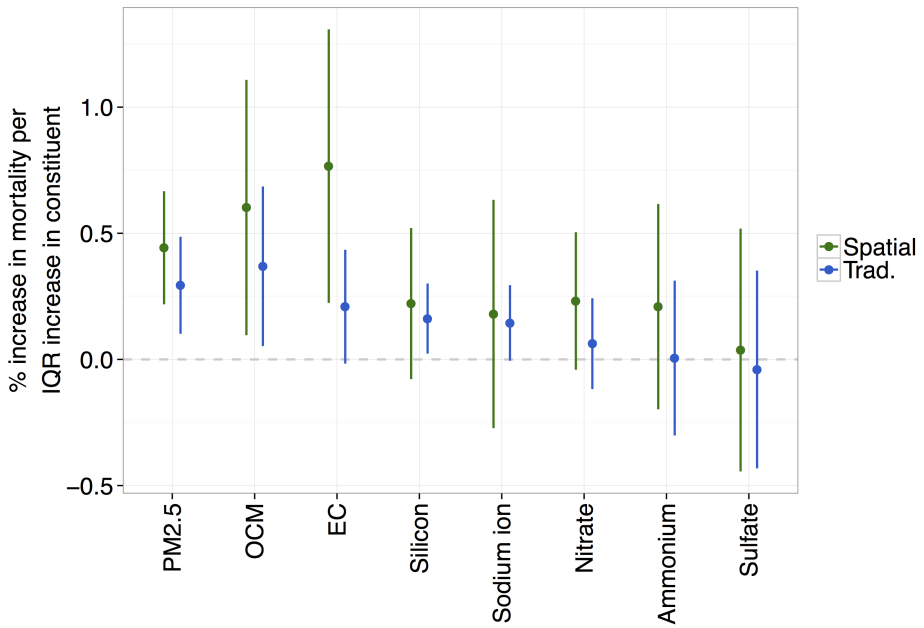

Figure 4.7: Estimated percent increase in mortality (95% PI) associated with an IQR increase in $PM_{2.5}$ mass and $PM_{2.5}$ constituents for previous-day exposure (lag 1).

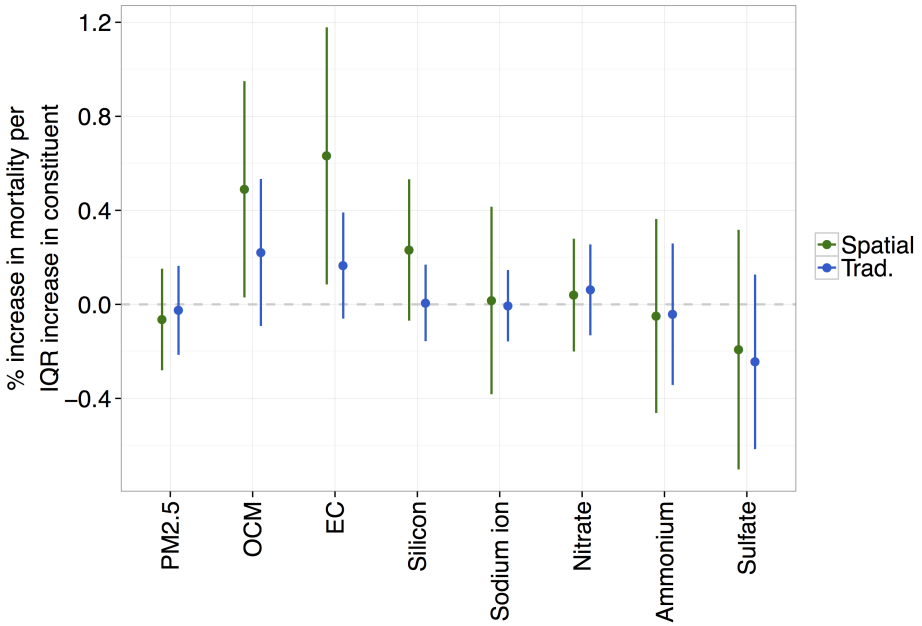

Figure 4.8: Estimated percent increase in mortality (95% PI) associated with an IQR increase in $PM_{2.5}$ mass and $PM_{2.5}$ constituents for exposure two days before (lag 2).

Table 4.3: Regression coefficients (ν_j from equation 4.6) comparing mortality risk estimates using estimated ambient average concentrations from the spatial model with the traditional approach.

	$PM_{2.5}$	OCM	EC	Silicon	Sodium ion	Nitrate	Ammonium	Sulfate
Lag 0	0.98	1.20	1.26	1.42	1.29	1.12	1.08	1.00
Lag 1	1.06	1.29	1.32	1.22	1.39	1.05	1.12	1.09
Lag 2	1.07	1.10	1.34	1.19	1.21	1.08	0.90	0.93

are not shown, but all had an absolute value of less than 0.02.

To determine whether constituent mortality effects identified from single pollutant models could be attributed to a smaller set of constituents, we fitted a multipollutant model with OCM, EC, silicon, and sodium ion as in Chapter 3. We fitted the multipollutant model for exposure on the previous-day and two days before, based on the evidence found at these lags in the single pollutant models. Figures 4.9 and 4.10 show multipollutant model results for the traditional approach and the spatial model respectively. Fewer observations were available to fit the multipollutant model than for the single pollutant models, and on average multipollutant models included 347 days of data (minimum = 56, maximum = 658). For the traditional approach at both lags, OCM and EC were attenuated in the multipollutant model compared with the single pollutant models. For an IQR increase in silicon on the previous day, we found a positive and statistically significant association with mortality for the traditional approach (0.18%, 95% PI: 0.04%, 0.33%). Comparing multipollutant and single pollutant models, associations for EC were less attenuated than associations for OCM for ambient averages estimated using the spatial model, however the standard errors for EC were larger for the multipollutant model compared with the single pollutant model.

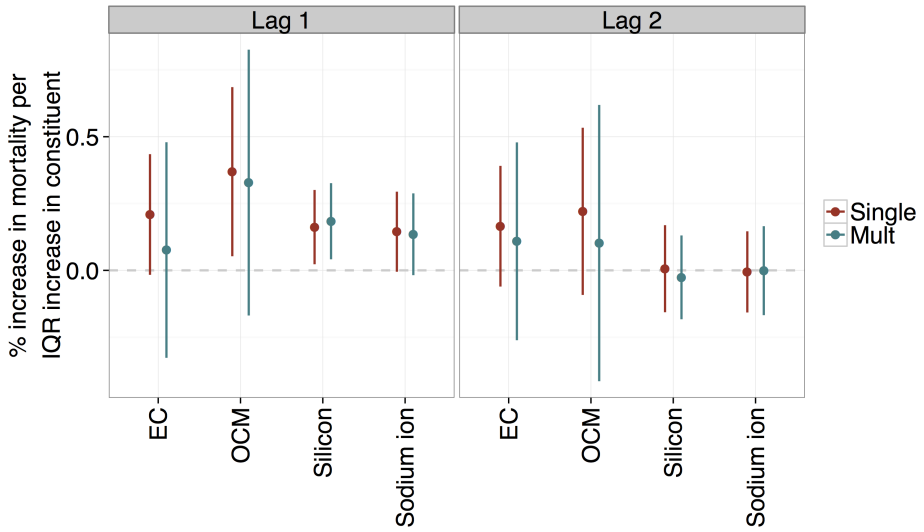

Figure 4.9: Estimated percent increase in mortality (95% PI) associated with IQR increases for a multipollutant model containing OCM, EC, silicon, and sodium ion, with ambient averages estimated using the traditional approach for lags 1 and 2.

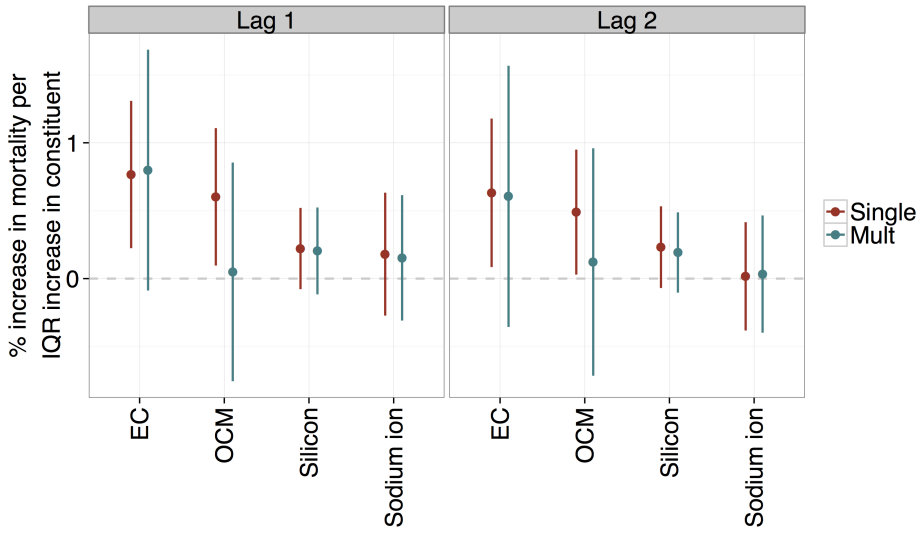

Figure 4.10: Estimated percent increase in mortality (95% PI) associated with IQR increases for a multipollutant model containing OCM, EC, silicon, and sodium ion for ambient averages estimated using the spatial model for lags 1 and 2.

4.5 Discussion

We assessed whether estimated associations between mortality and $PM_{2.5}$ total mass and $PM_{2.5}$ constituents differed between using a traditional approach and a spatial model for adjusting for spatial misalignment. Using both the traditional approach and the spatial model, we estimated national-level associations between mortality and $PM_{2.5}$ mass and seven major $PM_{2.5}$ constituents. For estimating ambient average pollutant concentrations using a spatial modeling approach, we fitted models separately for 6 regions in the US to account for potential differences in the correlation structure of pollutants between regions of the US with different sources of $PM_{2.5}$. Our proposed nonstationary model places minimal restrictions on the amount of data required and therefore can be fitted to the available $PM_{2.5}$ constituent data from the US EPA CSN. We found that estimated mortality effects were larger in magnitude using the spatial model to adjust for spatial misalignment compared with the traditional approach. For both the traditional approach and the spatial model, we found the largest associations between mortality and previous day exposure to OCM.

In Figure 4.4, we demonstrated that estimated ambient averages from the spatial model differ substantially from those using the traditional approach when the pollutant under consideration is spatially heterogeneous and only one central monitor is available. In such cases, the traditional approach may be inadequate because pollutant concentrations from one monitor may exhibit more variability over time than the true community-level ambient average. We compared estimated ambient averages between the two methods to adjust for spatial misalignment in Figure 4.5. The communities that show the largest differences potentially indicate areas where spatial

heterogeneity is large and a spatial model should be used to estimate ambient averages. Ultimately, we could not compare the accuracy of estimated ambient averages between the spatial model and the traditional approach because the true community-level averages are unknown and therefore differences shown in Figure 4.5 should be interpreted with some caution.

Previous work has also shown that the traditional approach may not be adequate for adjusting for spatial misalignment in studies of heterogeneous pollutants. In a simulation study, Strickland et al. (2013) found using one central monitor as the community-level ambient average increased bias, specifically attenuation of an estimated health effect. The bias was greater for spatially heterogeneous pollutants such as EC. Peng and Bell (2010) demonstrated that heterogeneous $PM_{2.5}$ constituents like EC were more strongly associated with hospital admissions using a spatial model. We could have considered other alternatives to our spatial model. Instead of fitting separate spatial models for each pollutant, Choi et al. (2009) developed a multivariate spatial-temporal model for $PM_{2.5}$ chemical constituents. Lee et al. (2009) used a spatial random effects model to account for spatial correlation for the association between respiratory hospital admissions and long-term exposure to $PM_{2.5}$ and NO_2 in Scotland. Yanosky et al. (2009) fitted a spatial-temporal model incorporating geographic information system data for estimating chronic exposure to $PM_{2.5}$, PM_{10}, and their difference ($PM_{10-2.5}$).

More complex nonstationary models have been proposed for geospatial data, however these methods have not been frequently applied to time series of air pollution and may require more data than are commonly available for $PM_{2.5}$ constituents (Higdon, 1998; Fuentes and Smith, 2001; Paciorek and Schervish, 2006). Fuentes and

Smith (2001) introduced a nonstationary model that is locally stationary, but parameter values can change across space more smoothly than our proposed nonstationary model. Higdon (1998) proposed a process-convolution approach, which applies a spatially and temporally varying convolution kernel to time series data. Paciorek and Schervish (2006) generalized Hidgon's approach by using a locally stationary model, but add a parameter that permits nonstationarity across the space. In future work, we could fit an anisotropic model (e.g. Luna and Genton (2005)) to $PM_{2.5}$ constituent data by incorporating a distance measure that allows for directional variation. An anisotropic model could improve estimation of $PM_{2.5}$ constituents because the correlation structure of highly variable constituents may vary with wind patterns. Many of the proposed nonstationary and anisotropic models could not be fitted to spatially and temporally sparse $PM_{2.5}$ constituent data. The approach we have taken here is simpler than methods proposed in the literature, but it can be applied to the available data and still accounts for spatial misalignment and nonstationarity.

4.5.1 Limitations

Since past work has shown that pollution concentrations and health outcomes are only weakly correlated (Peng and Bell, 2010), we did not use mortality data to estimate the ambient average from the spatial model. Determining spatial predictions using a fully Bayesian approach would require conditioning on mortality as well as the observed pollutant concentrations. We assumed that the fully Bayesian model is approximately equal to only conditioning on the observed pollutant concentrations. Let $\mathbf{m}$ be the observed pollutant concentrations and $\mathbf{Y}$ be the number of deaths. Conditional on the observed pollutant concentrations, we assumed the true ambient average $\mathbf{x}$ is independent of the health outcome, i.e. we assumed $P(\mathbf{x}|\mathbf{m},\mathbf{Y}) \approx P(\mathbf{x}|\mathbf{m})$.

This assumption would allow us to use the same estimated ambient average for a variety of health outcomes, instead of refitting the spatial model for each specific outcome of interest.

We have potentially underestimated the uncertainty from using a spatial model to adjust for spatial misalignment because we did not incorporate standard errors from estimating ambient averages in our health effects regression models. However, we do not report community-specific mortality risks, but rather pool risk estimates using a Bayesian hierarchical model that accounts for unexplained heterogeneity in mortality risks across communities. The hierarchical model produces standard errors for the national-level estimates that are generally robust to underestimation of community-specific mortality risk standard errors (Daniels et al., 2004). An expansion of the spatial model could incorporate methods to account for exposure measurement error that results from using estimated ambient average concentrations in our regression models. Similar methods have been previously developed for cohort data (Gryparis et al., 2009).

Neither approach used to adjust for spatial misalignment accounts for measurement error in observed pollutant concentrations that has been demonstrated in previous work (Bell et al., 2011). Depending on the type of measurement error, bias may exist in estimated ambient averages and also estimated health effects (Zeger et al., 2000). This paper did not address error from using ambient exposure data rather than personal exposure data, which are not available at the national level or for long timeframes (Dominici et al., 2000). However, work in cohort studies has indicated that improved exposure prediction does not always lead to better health effect estimation (Szpiro et al., 2011).

4.5.2 Conclusion

We demonstrated that estimated mortality effects of $PM_{2.5}$ mass and $PM_{2.5}$ constituents differed in magnitude between a traditional approach and a spatial model used to adjust for spatial misalignment. Associations with mortality estimated using the spatial model were generally larger in magnitude than associations estimated using the traditional approach, but had larger standard errors. We found evidence of associations between mortality and exposure to total mass $PM_{2.5}$, OCM, EC, and silicon. Both the traditional approach and the spatial model identified strongest associations between mortality and previous day OCM. For an IQR increase in previous-day exposure to OCM, we estimated a corresponding increase in mortality of 0.37% (95% PI: 0.05%, 0.69%) using the traditional approach and 0.60% (95% PI: 0.10%,1.11%) using the spatial model. We found that estimated health effects of $PM_{2.5}$ mass and $PM_{2.5}$ chemical constituents can depend on the method used to adjust for spatial misalignment.

Chapter 5

Censoring adjustment methods for source apportionment models

Sources of particulate matter (PM) air pollution are generally inferred from PM chemical constituent concentrations using source apportionment models. Concentrations of PM constituents are often censored below minimum detection limits and most source apportionment models cannot handle missing data. While methods used to impute or remove censored data have been evaluated for estimating summary statistics, it is not known how source apportionment methods, which are complex multivariate procedures, are affected by how censored data are treated. We demonstrated that when many data are censored, a likelihood-based imputation method leads to better source estimation compared with other methods used to adjust censored data. We compared our likelihood-based method to standard censoring adjustment methods when estimating sources in New York City. We found the estimated source distributions differed by both the censoring adjustment method and the choice of source apportionment method. We provide general guidance for adjusting censored PM constituent data in source apportionment, which is necessary for estimation of PM sources and their subsequent health effects.

5.1 Introduction

Particulate matter air pollution less than 2.5 μm in aerodynamic diameter ($PM_{2.5}$) is a complex chemical mixture (Bell et al., 2007) that originates from many different sources (Maykut et al., 2003; Ito et al., 2004; Hopke et al., 2006). $PM_{2.5}$ emitted from different sources likely varies in toxicity because $PM_{2.5}$ is a mixture of different chemical constituents that vary in toxicity (Ostro et al., 2007; Peng et al., 2009; Zhou et al., 2011; Krall et al., 2013). Determining which sources of $PM_{2.5}$ are most harmful to human health first requires good estimates of $PM_{2.5}$ emitted from different sources. Typically, $PM_{2.5}$ sources are not directly measured and must be inferred from daily $PM_{2.5}$ chemical constituent concentrations using source apportionment models. However, $PM_{2.5}$ constituent concentrations are frequently censored because the analytical methods used to obtain the data have a value below which they cannot determine whether a concentration is nonzero, referred to as the minimum detection limit (MDL) (Polissar et al., 2001; Kim et al., 2003; Maykut et al., 2003). Since many common source apportionment models cannot handle missing data, censored $PM_{2.5}$ constituent data must be imputed or removed before sources are estimated.

Previous environmental studies have quantified the impact of different methods to adjust censored concentrations, however the impact of these methods depends on what quantity is being estimated from the data (Helsel, 2010; Ganser and Hewett, 2010). Substituting censored concentrations with a constant between 0 and the MDL (e.g. $\frac{1}{2}\times$MDL) will lead to biased estimates of the mean and standard deviation (Helsel, 2006). For multivariate statistical procedures, such as principal component analysis (PCA), substitution methods may perform satisfactorily when data are censored. Farnham et al. (2002) applied PCA to groundwater chemicals and found that

substituting censored values with $\frac{1}{2}\times$MDL yielded principal component scores and loadings close to those obtained under no censoring. In a factor analysis setting, Aruga (1997) found that substituting censored values with a constant near zero lead to acceptable results, where acceptable was defined by the number of factors obtained, the variables associated with each factor, and the variance explained by each factor. Maximum likelihood methods have been used to impute censored data by estimating multivariate distributions (Hopke et al., 2001; Chen et al., 2013; Francis et al., 2009) and are generally preferred to substitution methods (Helsel, 2010). However, if the assumed distribution is incorrect, a likelihood-based approach may not yield good estimates of the censored data (Helsel, 2010).

Source apportionment methods are complex multivariate procedures, and therefore the method chosen to adjust censored data may impact source estimation differently than the estimation of summary statistics such as the mean. Most source apportionment models assume chemical constituent concentrations are related to $PM_{2.5}$ source concentrations using a modified factor analysis model. Factor analysis models attempt to recover latent variables (e.g. $PM_{2.5}$ sources) from the observed data. Source apportionment models differ from traditional factor analysis because they aim to estimate two non-negative matrices: the chemical contributions from each source, referred to as the source profiles, and the daily source concentrations. Common source apportionment methods include Absolute Principal Component Analysis (APCA) (Thurston and Spengler, 1985), Positive Matrix Factorization (PMF) (Paatero and Tapper, 1994; Norris et al., 2008), and Unmix (Henry, 1997; Norris et al., 2007). Some guidance exists on how to adjust censored data in source apportionment studies (Paatero and Hopke, 2003; Larson et al., 2004), however no studies have comprehensively examined how different adjustment methods impact source

apportionment models. Additionally, because different source apportionment models use different approaches to estimate sources, they may require different treatment of censored data.

In source apportionment studies, censored data are frequently substituted with a constant between 0 and the MDL for each constituent (commonly $\frac{1}{2}\times$MDL) (Larson et al., 2004; Marmur et al., 2005; Lee et al., 2008) or constituents with many censored observations are excluded or downweighted in the analysis (Kavouras et al., 2001; Querol et al., 2001; McDonald et al., 2003; Paatero and Hopke, 2003; Song et al., 2007). While likelihood-based methods to adjust censored data are preferred to substitution or exclusion methods (Helsel, 2010), likelihood-based approaches have not been evaluated in source apportionment studies and it is unknown whether a more complex censoring adjustment method will lead to better source attribution. Substitution and exclusion methods for adjusting censored data may be acceptable in source apportionment studies if the censored constituent is not critical for source estimation. However, if the censored constituent is necessary to distinguish similar sources, some censoring adjustment methods may limit our ability to correctly resolve $PM_{2.5}$ sources. For example, Figure 5.1 shows a time series of concentrations in New York City for two $PM_{2.5}$ constituents: aluminum, which has many concentrations that fall below the MDL, and calcium, which is completely observed. These constituents both contribute to a soil source of $PM_{2.5}$ in New York City (Ito et al., 2004), and estimation of this source may depend on the censoring adjustment method applied. However, if aluminum does not contribute to sources in New York City, the method chosen to adjust censored data may not impact source estimation.

This work offers two contributions. We first provided a comprehensive analysis

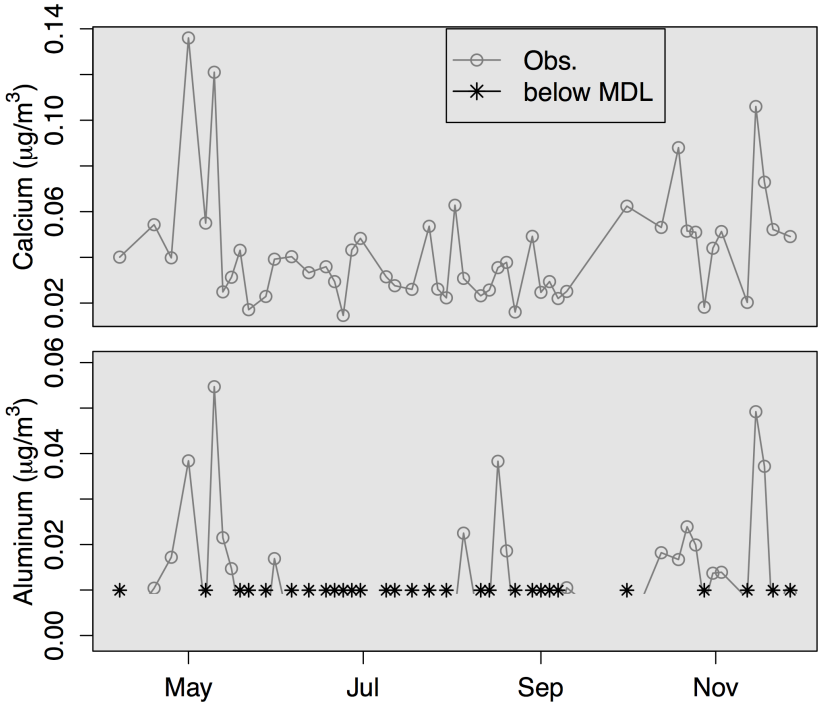

Figure 5.1: New York City time series for observed constituent data from April-November 2001 for two chemical constituents of $PM_{2.5}$: aluminum and calcium. Data below the MDL are marked with an asterisk at the MDL.

of the impact of different censoring adjustment methods on $PM_{2.5}$ source estimation. We examined two censoring adjustment methods that are commonly applied in source apportionment: substituting censored concentrations using $\frac{1}{2}\times$MDL and excluding or downweighting constituents with a large proportion of censored data. We developed a likelihood-based approach for imputing censored constituent concentrations that estimates the covariance between constituents and uses the estimated covariance to impute censored data. For both APCA and PMF, we demonstrated analytically and through simulation how each commonly-applied censoring adjustment method (substituting, excluding/downweighting) and our likelihood-based method impact source attribution. Second, we estimated sources in New York City and showed how source attribution varies by censoring adjustment method. Finally, we provided recommendations for adjusting censored PM chemical constituent data in source apportionment models. We detailed when a likelihood-based imputation method improves source attribution and when simpler methods might be acceptable. We have made software publicly available for imputing censored $PM_{2.5}$ constituent time series data using our likelihood-based approach (`http://bit.ly/14HZFBW`).

5.2 Methods

5.2.1 Source apportionment methods

Let $\mathbf{X}_{[n\times P]} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_P)$ be the available $PM_{2.5}$ chemical constituent data, where $\mathbf{x}_p$ is the vector of n daily concentrations for constituent p ($p = 1, ..., P$). Both APCA and PMF estimate two quantities from $\mathbf{X}$: the source concentration matrix $\mathbf{F}$ and the source profile matrix $\mathbf{\Lambda}$. The source concentration matrix $\mathbf{F}_{[n\times L]} = (\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_L)$, where $\mathbf{f}_l$ is n daily concentrations for source l. The source profile matrix $(\mathbf{\Lambda}_{[L\times P]})^T =$

$(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, ..., \boldsymbol{\lambda}_L)$, where $\boldsymbol{\lambda}_l$ is the profile for source l that gives the contributions of P chemical constituents to source l. We assume the number of sources L is known.

APCA

For each day t $(t = 1, ..., n)$ and constituent p, let $z_{tp} = \frac{x_{tp} - \bar{x}_p}{s_p}$ where $\bar{x}_p$ is the sample mean and s_p is the sample standard deviation for constituent p. Then, the vectors of length n that make up $\mathbf{Z}_{[n \times P]} = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_P)$ are all mean zero with unit variance. APCA first finds absolute principal component scores, $\mathbf{A}$, by rotating and rescaling results from Principal Component Analysis (PCA):

1. Find $\tilde{\mathbf{V}}$, the matrix of the first L eigenvectors of $\mathbf{Z}^T\mathbf{Z}$, using PCA, where the norm of each column of $\mathbf{V}$ is equal to the square root of the corresponding eigenvalue.

2. Find $\mathbf{V} = \tilde{\mathbf{V}}\mathbf{R}$, where $\mathbf{R}$ is the $L \times L$ varimax rotation matrix that satisfies

   $\mathbf{R} = \arg\max_{\mathbf{R}} \sum_{p=1}^{P} \sum_{l=1}^{L} (\tilde{\mathbf{V}}\mathbf{R})_{pl}^4 - \frac{1}{P} \sum_{l=1}^{L} (\sum_{p=1}^{P} (\tilde{\mathbf{V}}\mathbf{R})_{pl}^2)^2$ (Harris and Kaiser, 1964) and where P is the number of constituents

Then $\mathbf{A}_{[n \times L]} = (\mathbf{X}\mathbf{S}^{-1})[\text{Cor}(\mathbf{X})^{-1}\mathbf{V}]$, where $\mathbf{S} = \text{diag}(s_1, s_2, ..., s_P)$ is a diagonal matrix of the sample standard deviations of the columns of $\mathbf{X}$. The absolute principal component scores, $\mathbf{A}$, are the scaled but uncentered data rotated into the factor space. In the next step, total mass $PM_{2.5}$ is regressed on $\mathbf{A}$. The estimated coefficients, $\hat{\boldsymbol{\eta}}$, are then used to estimate source concentrations $f_{tl} = a_{tl} \times \hat{\eta}_l$. To find the source profiles, we regress daily concentrations for constituent p on the estimated source concentrations $\mathbf{F}$, $E[x_{tp} \mid \mathbf{F}] = \nu_p + \sum_{l=1}^{L} f_{tl} \times \lambda_{lp}$ to obtain the source profile matrix $\hat{\boldsymbol{\Lambda}}$. We implemented APCA using R version 3.0.2 (R Core Team, 2012).

Positive Matrix Factorization

Positive Matrix Factorization (PMF) finds $\mathbf{\Lambda}$ and $\mathbf{F}$ that minimize

$$\sum_{p=1}^{P}\sum_{t=1}^{T}\left(\frac{x_{tp}-\sum_{l=1}^{L} f_{tl}\times\lambda_{lp}}{u_{tp}}\right)^2 \tag{5.1}$$

subject to $\lambda_{lp}\geq 0$ and $f_{tl}\geq 0$ for all p,l,t. Uncertainties u_{tp} are selected to reflect the relative certainty about each x_{tp}. Since APCA assumes all uncertainties are equal, we set $u_{tp}=1$ for all t,p in PMF to make results more comparable between source apportionment methods. The multilinear engine is a program that finds $\mathbf{\Lambda}$ and $\mathbf{F}$ by minimizing the objective function for PMF (equation 5.1) using a conjugate gradient algorithm (Paatero, 1999). We used the ME version 2 (ME-2) software released with the user interface program PMF version 3.0, which is distributed by the US EPA.

5.2.2 Adjusting censored data below the MDL

We compared three methods for adjusting censored $PM_{2.5}$ constituent concentrations:

1. Substitute censored concentrations with $\frac{1}{2}\times$MDL
2. Exclude or downweight constituents with many missing concentrations
3. Impute censored concentrations using our proposed likelihood-based approach

We chose the constant $\frac{1}{2}\times$MDL because it yields better PCA results than using 0 or the MDL (Farnham et al., 2002). For APCA, method (2) is implemented by first excluding constituents with more than 25% daily concentrations below the MDL from the analysis and then substituting remaining censored concentrations with $\frac{1}{2}\times$MDL. For implementing method (2) for PMF, we substituted censored data with $\frac{1}{2}\times$MDL and then estimate the signal-to-noise ratio (SNR) for each constituent, as defined by

(Paatero and Hopke, 2003). If $0.2 < \text{SNR} < 2$, all uncertainties for the constituent are increased three-fold, and if $\text{SNR} \leq 0.2$, the constituent is excluded from the analysis. Both the $\frac{1}{2}\times$MDL approach (Song et al., 2001; Larson et al., 2004) and the exclude and/or downweight method (Rizzo and Scheff, 2007; Song et al., 2007) are frequently applied in the source apportionment literature.

Likelihood-based approach for imputing censored data

We developed a likelihood-based approach to multiply impute censored data. For each day t, we assumed $\log(\mathbf{x_t}) \sim \text{MVN}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, where $\mathbf{x}_t$ is the vector of P constituent concentrations on day t. We assumed the data are independent across time, which is an assumption of most source apportionment models and is likely reasonable in this application since $PM_{2.5}$ constituent monitors generally only sample $PM_{2.5}$ every sixth day.

We first estimated $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ using a Markov Chain Monte Carlo approach to sample from the posterior distributions of $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$, since censored data make using standard maximum likelihood estimators difficult. We used conjugate priors $\boldsymbol{\theta} \sim \text{MVN}(\mathbf{0}, 10^5\mathbf{I})$ and $\boldsymbol{\Sigma} \sim \text{inv-Wishart}(P+1, \mathbf{I})$, where $\mathbf{I}$ is the $P \times P$ identity matrix. We directly sampled from the posterior distributions of $\boldsymbol{\theta}$, $\boldsymbol{\Sigma}$, and the censored constituent concentrations using Gibbs sampling. Letting $\mathbf{Y} = \log(\mathbf{X})$, the full conditionals for $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ are

$$(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}, \mathbf{Y}) \sim \text{MVN}\left(\left(10^{-5}\mathbf{I} + n\boldsymbol{\Sigma}^{-1}\right)^{-1} \left(n\boldsymbol{\Sigma}^{-1}\bar{\mathbf{y}}\right), \left(10^{-5}\mathbf{I} + n\boldsymbol{\Sigma}^{-1}\right)^{-1} \right) \quad (5.2)$$

$$(\boldsymbol{\Sigma} \mid \boldsymbol{\theta}, \mathbf{Y}) \sim \text{inv-Wishart}\left(n + P + 1,\ \mathbf{I} + \sum_{t=1}^{n} (\mathbf{y}_t - \boldsymbol{\theta})(\mathbf{y}_t - \boldsymbol{\theta})^T \right) \quad (5.3)$$

where $\bar{\mathbf{y}} = (\bar{y}_1, \bar{y}_2, ..., \bar{y}_P)^T$. For each day t, let y_{tp} be the logged concentration for

a censored constituent p and $\mathbf{y}_{tq}$ be the logged concentrations for the remaining q constituents. The distribution of y_{tp} conditional on $\mathbf{y}_{tq}$ is truncated normal,

$$(y_{tp} \mid \mathbf{y}_{tq}, \boldsymbol{\theta}, \boldsymbol{\Sigma}) \sim \text{ trunc-N}\left(\theta_p + \boldsymbol{\Sigma}_{pq}\boldsymbol{\Sigma}_q^{-1}(\mathbf{y}_{tq} - \boldsymbol{\theta}_q),\ \Sigma_p - \boldsymbol{\Sigma}_{pq}\boldsymbol{\Sigma}_q^{-1}\boldsymbol{\Sigma}_{pq}^T\right) \tag{5.4}$$

where y_{tp} is truncated above by the log of its MDL, $\boldsymbol{\Sigma}_{pq}$ is the covariance between constituent p and the remaining constituents q, and θ_p, $\boldsymbol{\theta}_q$, Σ_p, and $\boldsymbol{\Sigma}_q$ refer to the subsets of $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ corresponding to constituents p and q.

We drew 50,000 samples from the joint distribution of $\boldsymbol{\theta}$, $\boldsymbol{\Sigma}$, and the censored data by iteratively sampling from the three distributions (equations (5.2), (5.3), and (5.4)) and updating the values for $\boldsymbol{\theta}$, $\boldsymbol{\Sigma}$, and each censored y_{tp}. Let $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}$ be the posterior means of $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ over the last 25,000 iterations. We imputed each censored concentration y_{tp} using a random draw from the truncated normal in equation 5.4, conditioning on observed constituents on day t and replacing $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ with $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}$. We created 10 imputed logged constituent datasets and then exponentiated each to obtain constituent concentrations on the original scale.

5.3 Impact of commonly applied censoring adjustment methods on source estimation

5.3.1 APCA

In this section, we showed how different censoring adjustment methods impact the estimation of $Cor(\mathbf{X})$ and then demonstrated how this impacts estimation of $PM_{2.5}$ sources using APCA. In order to estimate $PM_{2.5}$ sources using APCA, PCA is first applied to the centered and scaled data $\mathbf{Z}$. The first principal component (PC) satisfies $(\arg\max_{||\boldsymbol{\alpha}||=1} Var(\boldsymbol{Z}\boldsymbol{\alpha}))$. Since $Var(\boldsymbol{Z}\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T Var(\boldsymbol{Z})\boldsymbol{\alpha} = \boldsymbol{\alpha}^T Cor(\mathbf{X})\boldsymbol{\alpha}$, if

the sample correlation of the adjusted data is similar to the correlation for the uncensored, unobserved data, the first PC estimated from the adjusted data will be close to the first PC obtained from the uncensored data.

Commonly-applied censoring adjustment methods

Suppose concentrations below the MDL are substituted with a constant c_p. Let $\mathbf{W}$ be the imputed data such that $w_{tp} = c_p$ if $x_{tp} < m_{tp}$ and $w_{tp} = x_{tp}$ otherwise, where m_{tp} is the MDL corresponding to x_{tp}. Assume c_p is the sample mean of concentrations below the MDL, $\frac{1}{|B^{(p)}|}\sum_{t\in B^{(p)}} x_{tp}$, where $B^{(p)} = \{t : x_{tp} < m_{tp}\}$. Then the covariance between two substituted variables, $\mathbf{w}_1$ and $\mathbf{w}_2$, is

$$\begin{aligned}\widehat{Cov}(\mathbf{w}_1,\mathbf{w}_2) = \widehat{Cov}(\mathbf{x}_1,\mathbf{x}_2) - b^{(1)}\widehat{Cov}_{B^{(1)}}(\mathbf{x}_1,\mathbf{x}_2) \\ - b^{(2)}\widehat{Cov}_{B^{(2)}}(\mathbf{x}_1,\mathbf{x}_2) + b^{(1,2)}\widehat{Cov}_{B^{(1,2)}}(\mathbf{x}_1,\mathbf{x}_2) + \frac{|B^{(1,2)}|}{n-1}\zeta \qquad (5.5)\end{aligned}$$

where $B^{(1,2)} = \{t : x_{t1} < m_{t1},\ x_{t2} < m_{t2}\}$ and $b^{(\cdot)} = \frac{|B^{(\cdot)}|-1}{n-1}$. The terms $\widehat{Cov}_{B^{(\cdot)}}(\mathbf{x}_1,\mathbf{x}_2)$ are sample covariances restricted to the set $B^{(\cdot)}$. The last term

$$\zeta = \prod_{p=1}^{2}\left(\frac{1}{|B^{(1,2)}|}\sum_{t\in B^{(1,2)}} x_{tp} - c_p\right)$$

will generally be small since the differences are of two quantities that fall between 0 and the MDL. The covariance between the two substituted variables will be close to the true covariance, $\widehat{Cov}(\mathbf{w}_1,\mathbf{w}_2) \approx \widehat{Cov}(\mathbf{x}_1,\mathbf{x}_2)$, when

1. All restricted covariance terms $\widehat{Cov}_{B^{(\cdot)}}(\mathbf{x}_1,\mathbf{x}_2) \approx 0$ and $\zeta \approx 0$ or
2. All $b^{(\cdot)}$ are near zero

These conditions yield two situations when substituting censored constituents with a constant is sufficient: (a) constituents are uncorrelated with each other for values

below the MDL or (b) few data are censored. In general, we also would not expect $\widehat{Var}(\mathbf{w}_p) = \widehat{Cov}(\mathbf{w}_p, \mathbf{w}_p) \approx \widehat{Cov}(\mathbf{x}_p, \mathbf{x}_p) = \widehat{Var}(\mathbf{x}_p)$ and therefore $Cor(\mathbf{W})$ will be biased for $Cor(\mathbf{X})$.

If c_p is not equal to the sample mean of concentrations below the MDL, but is unbiased for observations below the MDL, the terms of the form $\widehat{Cov}_{B^{(\cdot)}}(\mathbf{x}_1, \mathbf{x}_2)$ will not be covariances or variances. If the substituted constant is not unbiased, then the expression for $Cov(\mathbf{w}_1, \mathbf{w}_2)$ is more complicated and farther from $Cov(\mathbf{x}_1, \mathbf{x}_2)$. In general, substituting data below the MDL using a constant will result in incorrect estimation of $Cor(\mathbf{X})$

In the extreme case, when all concentrations for a variable are substituted with a constant, then $\widehat{Var}(\mathbf{w}_p) = 0$ and $\widehat{Cov}(\mathbf{w}_p, \mathbf{w}_q) = 0$ for any q.

Likelihood-based approach

Suppose the logged constituent concentrations $\log(\mathbf{x}_t) = \mathbf{y}_t \sim$ trunc-MVN$(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, truncated above by the log of the MDLs. In the likelihood-based approach, we estimate $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ using the posterior means $\hat{\boldsymbol{\theta}} = (10^{-5}\mathbf{I} + n\boldsymbol{\Sigma}^{-1})^{-1}(n\boldsymbol{\Sigma}^{-1}\bar{\mathbf{y}})$ and $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\left(\mathbf{I} + \sum_{t=1}^{n}(\mathbf{y}_t - \boldsymbol{\theta})(\mathbf{y}_t - \boldsymbol{\theta})^T\right)$ where $\lim_{n\to\infty}\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ and $\lim_{n\to\infty}\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}$. The censored data $\mathbf{W}$ are imputed such that $\log(\mathbf{w}_t) \approx$ trunc-MVN$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}})$. Therefore, we developed the likelihood based approach to impute data so that the covariance of the imputed data is close to the covariance of the true, unobserved data.

Impact of poor variance estimation on source apportionment

For PCA, we can rewrite $Var(\mathbf{Z}\boldsymbol{\alpha})$ as

$$\begin{aligned} Var(\mathbf{Z}\boldsymbol{\alpha}) =& Var\left(\mathbf{Z}_{p_1}\boldsymbol{\alpha}_{p_1} + \mathbf{Z}_{p_2}\boldsymbol{\alpha}_{p_2}\right) \\ =& \boldsymbol{\alpha}_{p_1}^T Cor\left(\mathbf{X}_{p_1}\right)\boldsymbol{\alpha}_{p_1} + \boldsymbol{\alpha}_{p_2}^T Cor\left(\mathbf{X}_{p_2}\right)\boldsymbol{\alpha}_{p_2} + 2\boldsymbol{\alpha}_{p_1}^T Cor\left(\mathbf{X}_{p_1}, \mathbf{X}_{p_2}\right)\boldsymbol{\alpha}_{p_2} \end{aligned}$$

where $\mathbf{X}_{p_1}$ and $\boldsymbol{\alpha}_{p_1}$ represent the data and first PC for p_1 completely observed constituents and $\mathbf{X}_{p_2}$ and $\boldsymbol{\alpha}_{p_2}$ represent the data and first PC for p_2 constituents with censored data below the MDL. Because the first p_1 variables are observed, the correlation between these variables, $Cor\left(\mathbf{X}_{p_1}\right)$, will unaffected by censoring. However, the terms $Cor\left(\mathbf{X}_{p_2}\right)$ and $Cor\left(\mathbf{X}_{p_1}, \mathbf{X}_{p_2}\right)$ will be impacted by how the censored data is adjusted.

As shown in section 5.3.1, if we substitute censored concentrations with a constant, as in the $\frac{1}{2}\times$MDL method, then $Cor\left(\mathbf{X}_{p_2}\right)$, $Cor\left(\mathbf{X}_{p_1}, \mathbf{X}_{p_2}\right)$, and by extension $\boldsymbol{\alpha}$, will be poorly estimated. Excluding the p_2 censored variables completely, we effectively constrain $\boldsymbol{\alpha}_{p_2} = 0$ and elements of $|\boldsymbol{\alpha}_{p_1}|$ will likely be too large if the p_2 variables contribute to the first factor. We have developed a likelihood-based approach that aims to impute the data such that $Cor(\mathbf{W}) \approx Cor(\mathbf{X})$, which will hopefully lead to a better estimate of $\boldsymbol{\alpha}$.

The remaining PCs are obtained by solving

$$\boldsymbol{\alpha}_{l'} = \arg\max_{||\boldsymbol{\alpha}||=1} Var\left(\left(\mathbf{Z} - \mathbf{Z}\sum_{l=1}^{l'-1}\boldsymbol{\alpha}_l\boldsymbol{\alpha}_l^T\right)\boldsymbol{\alpha}\right) \tag{5.6}$$

Therefore, incorrect estimation of the first PC will lead to incorrect estimation of later PCs. Since the PCs are used in APCA to find both the source concentrations

and the source profiles, poor estimation of each successive PC will impact source apportionment results if the amount of censoring is substantial.

5.3.2 PMF

To demonstrate how censoring adjustment methods impact estimation of $\mathbf{F}$ and $\mathbf{\Lambda}$ in PMF, we use alternating least squares (ALS) to minimize equation 5.1. The ME-2 program uses a conjugate gradient algorithm for PMF because ALS convergence can be slow and is therefore not practical in application. ALS solves equations of the form $\mathbf{X} = \mathbf{F\Lambda}$ using the least squares solution, $\hat{\mathbf{\Lambda}} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{X}$. Assuming all uncertainties are equal to one, we can find PMF solutions by alternating between solving $\mathbf{X} = \mathbf{F\Lambda}$ for $\mathbf{\Lambda}$ and $\mathbf{X}^T = \mathbf{\Lambda}^T\mathbf{F}^T$ for $\mathbf{F}^T$.

Denote the imputed chemical constituent dataset by $\mathbf{W}$ such that $\mathbf{W} = \mathbf{X} + (\mathbf{W} - \mathbf{X})$. Then using $\mathbf{W}$ and an initial guess $\mathbf{F}_{(0)}$, we find

$$\begin{aligned}\tilde{\mathbf{\Lambda}}_{(1)} &= (\mathbf{F}_{(0)}^T\mathbf{F}_{(0)})^{-1}\mathbf{F}_{(0)}^T\mathbf{W} \\ &= (\mathbf{F}_{(0)}^T\mathbf{F}_{(0)})^{-1}\mathbf{F}_{(0)}^T\mathbf{X} + \boldsymbol{\delta} = \mathbf{\Lambda}_{(1)} + \boldsymbol{\delta}\end{aligned} \tag{5.7}$$

where $\mathbf{\Lambda}_{(1)}$ is the estimate of $\mathbf{\Lambda}$ we would obtain under no censoring and $\boldsymbol{\delta}$ is equal to $(\mathbf{F}_{(0)}^T\mathbf{F}_{(0)})^{-1}\mathbf{F}_{(0)}^T(\mathbf{W} - \mathbf{X})$, which will be near zero if $\mathbf{F}_{(0)}$ is uncorrelated with $(\mathbf{W} - \mathbf{X})$.

To obtain the estimate of the source concentrations $\mathbf{F}$ under the adjusted data ($\tilde{\mathbf{F}}_{(1)}$) ALS solves $\mathbf{W}^T = \tilde{\mathbf{\Lambda}}_{(1)}^T\mathbf{F}^T$ for $\mathbf{F}^T$. This regression has both error in the regressor, $\tilde{\mathbf{\Lambda}}_{(1)} = \mathbf{\Lambda}_{(1)} + \boldsymbol{\delta}$, and in the outcome, $\mathbf{W} = \mathbf{X} + (\mathbf{W} - \mathbf{X})$. If each entry in $\boldsymbol{\delta}$ is near zero and $\mathbf{\Lambda}_{(1)}$ is uncorrelated with $(\mathbf{W} - \mathbf{X})$, then $\tilde{\mathbf{F}}_{(1)}$ will be close to the estimate obtained under no censoring.

Since ALS alternates estimating $\mathbf{F}$ and $\mathbf{\Lambda}$ until convergence, the errors in each iteration propagate unless the deviations $\mathbf{W}-\mathbf{X}$ are uncorrelated with both $\mathbf{\Lambda}$ and $\mathbf{F}$ at each iteration. Substituting the censored data with a constant does not guarantee the deviations of the constant from the true unobserved concentrations will be uncorrelated with $\mathbf{\Lambda}$ or $\mathbf{F}$. Excluding a constituent that contributes to a source will likely decrease the estimated source concentration across time, since PMF estimates the concentrations of the source using contributing constituent concentrations. Additionally, excluding constituents will provide incorrect source profiles by failing to identify that constituent as a contributing pollutant to a source.

When PMF includes uncertainties, we can solve for $\mathbf{F}$ and $\mathbf{\Lambda}$ using alternating weighted least squares, where the weights correspond to the uncertainties. For the source concentrations, we solve $\tilde{\mathbf{F}}^T_{(1)} = (\tilde{\mathbf{\Lambda}}_{(1)}\mathbf{U}\tilde{\mathbf{\Lambda}}^T_{(1)})^{-1}\tilde{\mathbf{\Lambda}}_{(1)}\mathbf{U}\mathbf{W}^T$, where $\mathbf{U} = \text{diag}(u_1, u_2, ..., u_P)$ is the matrix of uncertainties for each constituent. Using this expression, concentrations of particular source will be underestimated if constituents that contribute to that source are downweighted. In the downweight/exclude censoring adjustment method for PMF, constituent uncertainties are the same across observations and therefore computation of $\tilde{\mathbf{\Lambda}}_{(1)}$ is the same as in equation 5.7.

Our proposed likelihood-based approach for imputing censored data does not impute censored data so that the deviations $\mathbf{W}-\mathbf{X}$ are uncorrelated with $\mathbf{\Lambda}$ and $\mathbf{F}$. For example, if the constituents that have large deviations $\mathbf{w}_p - \mathbf{x}_p$ also contribute a lot to certain sources (e.g. λ_{pl} is large for some source l), we would not necessarily expect the likelihood-based method to perform well. When PMF is used for source apportionment, a likelihood-based approach for imputing censored data may not have an advantage over other, commonly applied censoring adjustment methods.

5.4 Simulation study

We have shown analytically that different censoring adjustment methods will perform differently depending both on the source apportionment model and the amount of censoring. A simulation study comparing how different censoring adjustment methods affect $PM_{2.5}$ source estimation will demonstrate the practical benefits and disadvantages of selecting specific censoring adjustment methods. We conducted a simulation study to compare three methods for adjusting censored data: (1) $\frac{1}{2} \times$MDL (2) excluding and/or downweighting constituents and (3) our likelihood-based approach.

5.4.1 SPECIATE database for source profiles

In our simulation study, we incorporated data from the US EPA SPECIATE database (version 4.2), which contains $PM_{2.5}$ source profiles collected throughout the US for 53 chemical constituents. We cleaned SPECIATE such that (1) there were $P = 23$ $PM_{2.5}$ chemical constituents (Table 5.1), (2) the source profiles were normalized to represent the percent contribution of each chemical constituent to a source and (3) the sources fell into one of 7 major source categories in the US: wood burning (wood), diesel exhaust (diesel), road dust (dust), motor vehicles (vehicle), coal combustion (coal), oil combustion, and metals production (details in the Supplementary material, section 5.7). While the 23 constituents are a fraction of the over 50 constituents that make up $PM_{2.5}$ (Bell et al., 2007), generally source apportionment focuses on a subset of 20-30 constituents, which include constituents that contribute most to $PM_{2.5}$ by mass (e.g., sulfate, nitrate, organic carbon) (Bell et al., 2007), smaller constituents previously identified as toxic (e.g., arsenic, nickel, vanadium, selenium) (Ito et al.,

Table 5.1: The 23 $PM_{2.5}$ chemical constituents in the cleaned SPECIATE database and used in the simulation study.

Aluminum	Arsenic	Bromine	Calcium	Chlorine	Copper
Elemental Carbon	Iron	Potassium	Manganese	Sodium	Nickel
Nitrate	Organic Carbon	Phosphorus	Lead	Selenium	Silicon
Sulfate	Strontium	Titanium	Vanadium	Zinc	

2004; Franklin et al., 2008; Bell et al., 2009; Zanobetti et al., 2009) and key contributors to common sources (e.g. calcium, aluminum, titanium) (Ito et al., 2004; Nikolov et al., 2007).

5.4.2 Simulating $PM_{2.5}$ constituent data

To simulate $PM_{2.5}$ chemical constituent data $\mathbf{X}$, we used the Schur (element-wise) product of lognormal errors $\boldsymbol{e}$ with the product of a source concentration matrix, $\mathbf{F}$, and a source profile matrix, $\mathbf{\Lambda}$,

$$\mathbf{X}_{[n\times P]} = (\mathbf{F}_{[n\times L]}\mathbf{\Lambda}_{[L\times P]}) \circ \boldsymbol{e}_{[n\times P]} \tag{5.8}$$

where $\log(e_{tp}) \overset{IID}{\sim} N(0, 0.01^2)$. We used lognormally distributed errors to ensure constituent concentrations were non-negative. For the rows of $\mathbf{\Lambda}_{[L\times P]}$, we selected prototypical source profiles from the cleaned SPECIATE database for wood, diesel, dust, vehicle, and coal to represent two hypothetical communities with different numbers of sources: $L = 3$ sources (wood/ diesel/ dust) and $L = 5$ sources (wood/ diesel/ dust/ vehicle/ coal). To generate the concentration time series $\mathbf{f}_l$ for each source l, we generated $n = 1000$ independent lognormal concentrations with means and standard deviations in Table 5.2, chosen to approximately reflect the distribution of sources reported in the literature (Ito et al., 2004; Lingwall et al., 2008).

For both the 3-source and 5-source scenarios, we simulated 300 $\mathbf{F}$ datasets and created 300 datasets $\mathbf{X}$. To introduce different degrees of censoring, we first randomly

Table 5.2: Means and standard deviations of the lognormal distribution for each source in the simulation study.

source	mean ($\mu g/m^3$)	standard deviation ($\mu g/m^3$)
wood	2.56	1.31
diesel	5.87	0.75
dust	5.14	0.82
vehicle	1.73	4.33
coal	7.41	0.63

selected 2 or 11 of the 23 constituents to be censored. Then we created five censored datasets for each $\mathbf{X}$ by censoring these randomly selected constituents at their 20%, 50%, or 80% quantiles from the observed constituent concentration distributions. We did not censor 11 constituents at 80% because nearly 40% of the total data would be censored and source apportionment is not practical in this setting. These censoring scenarios yielded between 1.7% and 23.9% censored data across all constituents. We adjusted each censored dataset using three censoring adjustment methods. When constituents were censored at 20%, the exclude method used in APCA does not drop any constituents and therefore does not differ from the $\frac{1}{2}\times$MDL method. The different simulation scenarios are shown in Table 5.3.

5.4.3 Comparing source apportionment results between censoring adjustment methods

To compare the relative performance of different censoring adjustment methods in recovering $PM_{2.5}$ sources, we compared sources estimated using the uncensored, simulated data $\mathbf{X}$ with sources estimated from censored and then adjusted data $\mathbf{X}$. Specifically, we computed (a) the number of incorrectly identified sources under censoring, (b) the source concentration bias due to censoring, and (c) the ratio of source concentration sample variances between data with and without censoring. These measures

Table 5.3: Simulation study comparing censoring adjustment methods.

Source scenario	3 sources (wood/diesel/dust)
	5 sources (wood/diesel/dust/vehicle/coal)
Source apportionment method	APCA
	PMF
Censoring adjustment method	$\frac{1}{2}\times$MDL
	Likelihood-based
	Exclude and/or downweight constituents
Number (out of 23 total constituents) and quantile of constituents censored	
	2 constituents at $\{20\%, 50\%, \text{or } 80\%\}$
	11 constituents at $\{20\%, \text{or } 50\%\}$

reflect information about sources that is commonly reported in the source apportionment literature (Ito et al., 2004; Hopke et al., 2006) and capture both the identification of the source (a) and distributional properties of the source concentrations (b and c).

Sources are frequently identified by linking constituents that have large values in estimated source profiles to expert knowledge about the chemical makeup of different sources (Ito et al., 2004; Hopke et al., 2006). However in a large simulation study, individually inspecting each profile would be impossible. To identify sources, we used a k-nearest neighbors classification using our cleaned SPECIATE dataset. Details about the classification method can be found in the Supplementary material (Section 5.7). Source misclassification under censoring acts as a measure of misattribution (e.g. naming vehicle as diesel), which is important since a major aim of source apportionment is to identify sources of $PM_{2.5}$ in a community.

We used the classification method to identify sources obtained from the censored

data, which allowed us to compare source concentration means and variances between data with and without censoring. Although we ensured that we always correctly classified sources from the uncensored data, there were cases where the sources classified from the censored data did not match the uncensored sources. In these cases, we were unable to compare source means and variances between data with and without censoring. As a measure of source concentration bias due to censoring, we compared average difference in source concentrations means for source l, d_l, between data with and without censoring by reporting $\left(\sqrt{\frac{1}{L}\sum_{l=1}^{L} d_l^2}\right)$ across sources. For an estimated source l and simulated dataset i, let $\tilde{s}_{il}^2$ be the sample temporal source concentration variance from censored data and s_{il}^2 be the corresponding quantity for the uncensored data. To compare sample source concentration variances for each source and each amount of censoring, we computed $r_l = \exp\left\{\frac{1}{300}\sum_{i=1}^{300}\log\left(\frac{\tilde{s}_{il}^2}{s_{il}^2}\right)\right\}$ across 300 simulated datasets. Since the aim of this work was to quantify bias due to censoring and not to evaluate source apportionment methods, we did not compare estimated source concentration means and variances to the true means and variances used to generate the source concentrations.

For our likelihood-based method, we computed each quantity (a-c) for each of the 10 imputed datasets and took the 20% trimmed mean. This allowed us to discard the impact of datasets with extremely large imputed concentrations, which sometimes occurs with lognormally imputed data.

5.4.4 Results for source estimation

Across all measures and censoring adjustment methods, the impact of censoring on source attribution increased with the amount of censored data. We obtained the average number of sources misclassified of 3 or 5 total sources across simulations

using our k-nearest neighbors classification method. For APCA, while both the $\frac{1}{2}\times$MDL method and the likelihood-based method never misclassified sources, when constituents were excluded from the analysis, sources were frequently misclassified (Table 5.4). For example in the 3-source scenario using the exclude method, 1 of the 3 sources was incorrectly classified on average when more than 20% of the data were censored. Diesel was often misclassified, though wood and dust were sometimes misclassified as well. Under wood/ diesel/ dust/ vehicle/ coal, the most frequent misclassifications were wood, diesel, and vehicle, while dust was hardly ever misclassified.

Table 5.4: Average number of sources misclassified under different amounts of censoring for two source scenarios: 3 sources (wood/ diesel/ dust) and 5 sources (wood/ diesel/ dust/ vehicle/ coal). APCA was used for source apportionment.

Method	Sources	20%	50%	80%	20%	50%
Likelihood	3 sources	0.00	0.00	0.00	0.00	0.00
1/2MDL	3 sources	0.00	0.00	0.00	0.00	0.00
Exclude	3 sources	-	1.00	1.00	-	1.00
Likelihood	5 sources	0.00	0.00	0.00	0.00	0.00
1/2MDL	5 sources	0.00	0.00	0.00	0.00	0.00
Exclude	5 sources	-	1.00	0.00	-	2.00

Table 5.5 shows the average source concentration bias, d_l, squared and summed over sources $\left(\sqrt{\frac{1}{L}\sum_{l=1}^{L} d_l^2}\right)$ for APCA. Compared to other censoring adjustment methods, the likelihood-based method generally led to less source concentration bias due to censoring. When there were 3 sources, excluding constituents that were missing more than 25% daily concentrations seemed to decrease concentration bias due to censoring compared with the $\frac{1}{2}\times$MDL approach. However, excluding constituents did not perform uniformly better than $\frac{1}{2}\times$MDL when there were 5 sources.

Table 5.5: Average source concentration bias, d_l, squared and summed over sources $\left(\sqrt{\frac{1}{L}\sum_{l=1}^{L} d_l^2}\right)$ for two source scenarios: 3 sources (wood/ diesel/ dust) and 5 sources (wood/ diesel/ dust/ vehicle/ coal). APCA was used for source apportionment.

		2 constituents			11 constituents	
Method	Sources	20%	50%	80%	20%	50%
Likelihood	3 sources	0.00	0.00	0.02	0.02	0.23
1/2MDL	3 sources	0.11	0.19	0.09	0.56	0.87
Exclude	3 sources	-	0.05	0.06	-	0.12
Likelihood	5 sources	0.00	0.01	0.04	0.13	0.53
1/2MDL	5 sources	0.10	0.08	0.07	1.05	2.96
Exclude	5 sources	-	0.07	0.10	-	2.55

The estimated variance ratios between sources estimated with and without censored data for APCA are shown in Tables 5.6 and 5.7. Across censoring adjustment methods, some sources had overestimated variances under censoring and other sources had underestimated variances. For example, the variance of diesel was severely overestimated and the variance of vehicle was underestimated when many data were censored in the 5-source scenario (Table 5.7). Substituting censored data with $\frac{1}{2}\times$MDL led to good estimates of the source variances when few data were censored. The exclude method generally performed worse than imputing data with $\frac{1}{2}\times$MDL. Compared with other censoring adjustment methods, the likelihood-based approach led to sample temporal source variances under censoring closer to those estimated from uncensored data for APCA.

Results from PMF are included in Tables 5.8-5.11. Censored data made source classification difficult under PMF, but PMF frequently led to estimates of the source means that were less affected by censored data compared with APCA. Using PMF, the likelihood-based method frequently provided less or equally biased estimates of

Table 5.6: Exponentiated average log ratio of concentration variances between data with and without censoring for each source l, r_l, for sources wood/diesel/dust. APCA was used for source apportionment.

		2 constituents			11 constituents	
Method	source	20%	50%	80%	20%	50%
Likelihood	wood	1.00	1.00	1.00	1.00	1.00
1/2 MDL		1.00	1.00	1.00	1.00	1.00
Exclude		-	1.00	1.01	-	1.02
Likelihood	diesel	1.00	1.00	1.00	1.00	1.00
1/2 MDL		1.00	1.00	1.00	0.99	0.98
Exclude		-	0.99	0.99	-	0.97
Likelihood	dust	1.00	1.00	1.00	1.00	1.00
1/2 MDL		1.00	1.00	1.00	0.99	0.98
Exclude		-	0.98	0.98	-	0.96

Table 5.7: Exponentiated average log ratio of concentration variances between data with and without censoring for each source l, r_l, for sources wood/diesel/dust/vehicle/coal. APCA was used for source apportionment.

		2 constituents			11 constituents	
Method	source	20%	50%	80%	20%	50%
Likelihood	wood	1.00	1.00	1.00	1.00	1.00
1/2 MDL		1.00	1.00	1.00	1.00	1.00
Exclude		-	1.02	1.02	-	1.16
Likelihood	diesel	1.00	1.00	1.01	1.01	2.52
1/2 MDL		1.01	1.02	1.05	8.45	12.82
Exclude		-	1.10	1.22	-	82.38
Likelihood	dust	1.00	1.00	1.00	0.99	1.00
1/2 MDL		1.00	1.00	1.01	0.98	1.01
Exclude		-	1.05	1.06	-	1.68
Likelihood	vehicle	1.00	1.00	1.00	1.00	0.96
1/2 MDL		1.00	1.00	1.00	0.96	0.49
Exclude		-	0.99	0.99	-	0.62
Likelihood	coal	1.00	1.00	1.00	1.00	1.00
1/2 MDL		1.00	1.00	1.00	1.00	0.98
Exclude		-	1.00	0.99	-	0.69

Table 5.8: Average number of sources misclassified under different amounts of censoring for two source scenarios: 3 sources (wood/ diesel/ dust) and 5 sources (wood/ diesel/ dust/ vehicle/ coal). PMF was used for source apportionment.

		2 constituents			11 constituents	
Method	sources	20%	50%	80%	20%	50%
Likelihood	3 sources	0.00	0.00	0.17	0.00	0.67
1/2MDL	3 sources	0.00	0.00	0.00	1.00	1.00
Exclude/downweight	3 sources	0.00	0.00	0.00	1.00	1.00
Likelihood	5 sources	0.17	0.17	0.17	0.67	0.17
1/2MDL	5 sources	0.00	1.00	0.00	1.00	0.00
Exclude/downweight	5 sources	0.00	1.00	0.00	1.00	0.00

Table 5.9: Average source concentration bias, d_l, squared and summed over sources $\left(\sqrt{\frac{1}{L}\sum_{l=1}^{L} d_l^2}\right)$ for two source scenarios: 3 sources (wood/ diesel/ dust) and 5 sources (wood/ diesel/ dust/ vehicle/ coal). PMF was used for source apportionment.

		2 constituents			11 constituents	
Method	sources	20%	50%	80%	20%	50%
Likelihood	3 sources	0.00	0.00	0.00	0.03	0.45
1/2MDL	3 sources	0.00	0.01	0.00	0.19	0.48
Exclude/downweight	3 sources	0.00	0.02	0.00	0.20	0.50
Likelihood	5 sources	0.02	0.01	0.02	0.37	0.54
1/2MDL	5 sources	0.01	0.01	0.02	0.44	0.52
Exclude/downweight	5 sources	0.01	0.02	0.02	0.45	0.89

source means and variances compared with other commonly applied censoring adjustment methods. However, the relative performance of censoring adjustment methods was less consistent than when APCA was used for source apportionment.

5.4.5 Sensitivity analysis

We found the performances of censoring adjustment methods were consistent across an array of sensitivity analyses. For APCA, the likelihood-based method still performed best under a different combination of sources (Tables 5.12-5.15). Our results were robust to the choice of source concentration means and standard deviations and

Table 5.10: Exponentiated average log ratio of concentration variances between data with and without censoring for each source l, r_l, for sources wood/diesel/dust. PMF was used for source apportionment.

		2 constituents			11 constituents	
Method	source	20%	50%	80%	20%	50%
Likelihood	wood	1.00	1.00	1.00	0.99	0.96
1/2 MDL		1.00	1.00	1.00	0.94	0.95
Exclude/downweight		1.00	0.99	1.00	0.93	0.87
Likelihood	diesel	1.00	1.00	1.00	1.04	1.29
1/2 MDL		1.00	1.00	1.00	1.18	1.45
Exclude/downweight		1.00	1.00	1.00	1.19	1.36
Likelihood	dust	1.00	1.00	1.00	1.00	1.20
1/2 MDL		1.00	1.01	1.00	1.08	1.19
Exclude/downweight		1.00	1.00	1.00	1.08	1.10

Table 5.11: Exponentiated average log ratio of concentration variances between data with and without censoring for each source l, r_l, for sources wood/diesel/dust/vehicle/coal . PMF was used for source apportionment.

		2 constituents			11 constituents	
Method	source	20%	50%	80%	20%	50%
Likelihood	wood	1.00	1.00	1.00	0.94	0.94
1/2 MDL		1.00	1.00	1.00	0.89	1.02
Exclude/downweight		1.00	1.00	1.00	0.90	0.92
Likelihood	diesel	1.01	1.10	1.15	1.19	1.72
1/2 MDL		1.01	1.07	1.06	1.27	1.76
Exclude/downweight		1.02	1.06	1.03	1.28	1.82
Likelihood	dust	1.00	1.00	1.00	0.92	1.02
1/2 MDL		1.00	1.00	1.00	0.94	1.10
Exclude/downweight		1.00	1.00	1.00	0.96	1.06
Likelihood	vehicle	1.00	1.00	1.00	1.01	0.99
1/2 MDL		1.00	1.00	1.00	1.03	1.02
Exclude/downweight		1.00	1.00	1.00	1.02	0.99
Likelihood	coal	1.00	1.00	1.01	1.03	1.12
1/2 MDL		1.00	1.00	1.00	1.09	1.19
Exclude/downweight		1.00	1.00	1.00	1.09	1.30

Table 5.12: Means and standard deviations of the lognormal distribution used for sources dust/vehicle/diesel in the sensitivity analysis for the simulation study.

source	mean	standard deviation
dust	2.56	1.31
vehicle	5.87	0.75
diesel	5.14	0.82

Table 5.13: Average number of sources misclassified under different amounts of censoring for sources dust/vehicle/diesel. APCA was used for source apportionment.

	11 cons.			23 cons.	
Method	20%	50%	80%	20%	50%
Likelihood	0.00	0.00	0.00	0.00	0.00
1/2MDL	0.00	0.00	0.00	0.00	0.00
Exclude	-	0.00	0.00	-	1.00

the choice of prototypical source profiles from SPECIATE. As a sensitivity analysis, we also generated the source concentration data with an arbitrary covariance between the sources, but found results were similar to assuming sources were independent.

The simulated constituent data were not lognormally distributed, since the data were generated using the Schur product of lognormal errors and a linear combination of the lognormally distributed sources and fixed source profiles (equation 5.8). If the

Table 5.14: Average source concentration bias, d_l, squared and summed over sources $\left(\sqrt{\frac{1}{L}\sum_{l=1}^{L} d_l^2}\right)$ for sources dust/vehicle/diesel. APCA was used for source apportionment.

	2 constituents			11 constituents	
Method	20%	50%	80%	20%	50%
Likelihood	0.00	0.00	0.02	0.05	0.31
1/2MDL	0.08	0.12	0.08	0.51	0.84
Exclude	-	0.04	0.05	-	0.41

Table 5.15: Exponentiated average log ratio of concentration variances between data with and without censoring for each source l, r_l, for sources dust/vehicle/diesel. APCA was used for source apportionment.

		2 constituents			11 constituents	
Method	source	20%	50%	80%	20%	50%
Likelihood	dust	1.00	1.00	1.00	1.00	1.00
1/2 MDL		1.00	1.00	1.00	0.99	0.99
Exclude		-	1.00	1.00	-	0.98
Likelihood	vehicle	1.00	1.00	1.00	1.00	1.00
1/2 MDL		1.00	1.00	0.99	1.00	0.99
Exclude		-	0.99	0.99	-	0.97
Likelihood	diesel	1.00	1.00	1.00	1.00	1.00
1/2 MDL		1.00	1.00	1.00	0.99	0.99
Exclude		-	0.99	0.99	-	1.01

simulated constituent data were distributed as lognormal by design, our likelihood-based approach may have overperformed because it assumes a lognormal distribution. We tested whether the likelihood-based method was overperforming because of the lognormally distributed errors by generating $\mathbf{X}$ as in equation 5.8 using Gamma distributed errors $\boldsymbol{e}$ (shape=rate=25). We found results using Gamma distributed errors, in particular the results from our likelihood-based model, were unchanged.

5.5 $PM_{2.5}$ sources in New York City

5.5.1 Data

For estimating $PM_{2.5}$ sources in New York City, we used the Queens College monitor from the New York State Department of Environmental Conservation (`http://bit.ly/198xZcm`, accessed 29 February 2012) to create a dataset of 174 days from April 2001-December 2002 to match a previous analysis (Ito et al., 2004). Our dataset consisted of all constituents from the simulation study except sulfate (Table 5.1) as well as barium, cadmium, chromium, cobalt, magnesium, molybdenum,

sulfur, and ammonium ion, for a total of 30 constituents (Ito et al., 2004). Because some daily values for the MDL were missing, we used each constituent's maximum MDL, though results were similar to using the minimum MDL for each constituent. Using a constant MDL for each constituent over time also avoids creating false associations between variables that may be introduced when using a substitution method (Helsel, 2010). For the likelihood-based approach, we used 10 draws from the truncated lognormal distribution (equation 5.4) and used the 10% trimmed mean across imputations to create the concentration time series for each source.

Commonly, the methods used to obtain $PM_{2.5}$ chemical constituent concentrations from ambient monitors report daily concentrations that fall below the MDL, which may have no relationship to the true concentrations (Helsel, 2005a). In this data analysis, we compare source apportionment results between the three censoring adjustment methods used in the simulation study and results using the reported data below the MDL.

5.5.2 Results

In our New York City dataset, 15 of the constituents had less than 25% censored concentrations, 2 constituents had 25%- 50% censored concentrations, and 13 constituents had more than 50% concentrations below the MDL. We matched our source apportionment results to 4 $PM_{2.5}$ sources (soil, secondary sulfate, traffic, and residual oil/incineration) as reported by (Ito et al., 2004). Using APCA for source apportionment, the concentration means and standard deviations of the sources are similar in magnitude across censoring adjustment methods, with some differences (Table 5.16). For example, the estimated mean concentration of secondary sulfate ranges from 4.53

Table 5.16: Mean concentrations (standard deviations) in $\mu g/m^3$ for four sources in New York City estimated using APCA including soil, secondary sulfate (sec. SO_4^{-2}), traffic, and residual oil/incineration. Results using four different methods for adjusting censored data are shown: Reported data, Likelihood, $\frac{1}{2}\times$MDL, Exclude.

	Soil	Sec. SO_4^{-2}	Traffic	Res. oil/incineration
Reported	2.51 (2.59)	7.23 (6.86)	5.01 (5.35)	1.26 (1.82)
Likelihood	2.59 (1.82)	4.53 (5.41)	7.21 (5.65)	3.76 (3.31)
1/2 MDL	1.78 (2.19)	5.88 (6.74)	4.73 (6.09)	0.4 (1.44)
Exclude	2.56 (2.14)	5.86 (6.22)	4.28 (6.63)	1.68 (2.38)

$\mu g/m^3$ when constituents are excluded to 7.23 $\mu g/m^3$ using the reported data. Figure 5.2 shows APCA-estimated source time series from April 2001-September 2001 for the reported data and two censoring adjustment methods. The grey band shows the interquartile range of estimated time series using the likelihood-based method and demonstrates uncertainty in estimating source concentrations driven by censored data. The time series plots show similar trends across censoring adjustment methods but there exist some differences in the average source concentration and source concentration variability. Note that Figure 5.2 only displays data for 48 of 174 days of data in New York City so that the points can be seen clearly and therefore does not exactly reflect the means and standard deviations reported in Table 5.16.

We also applied PMF to data from New York City and found the source estimates were similar across censoring adjustment methods (Table 5.17). Differences between our results and those reported for New York City by (Ito et al., 2004) may be due to a different dataset, different implementations of source apportionment models, and different methods of source identification.

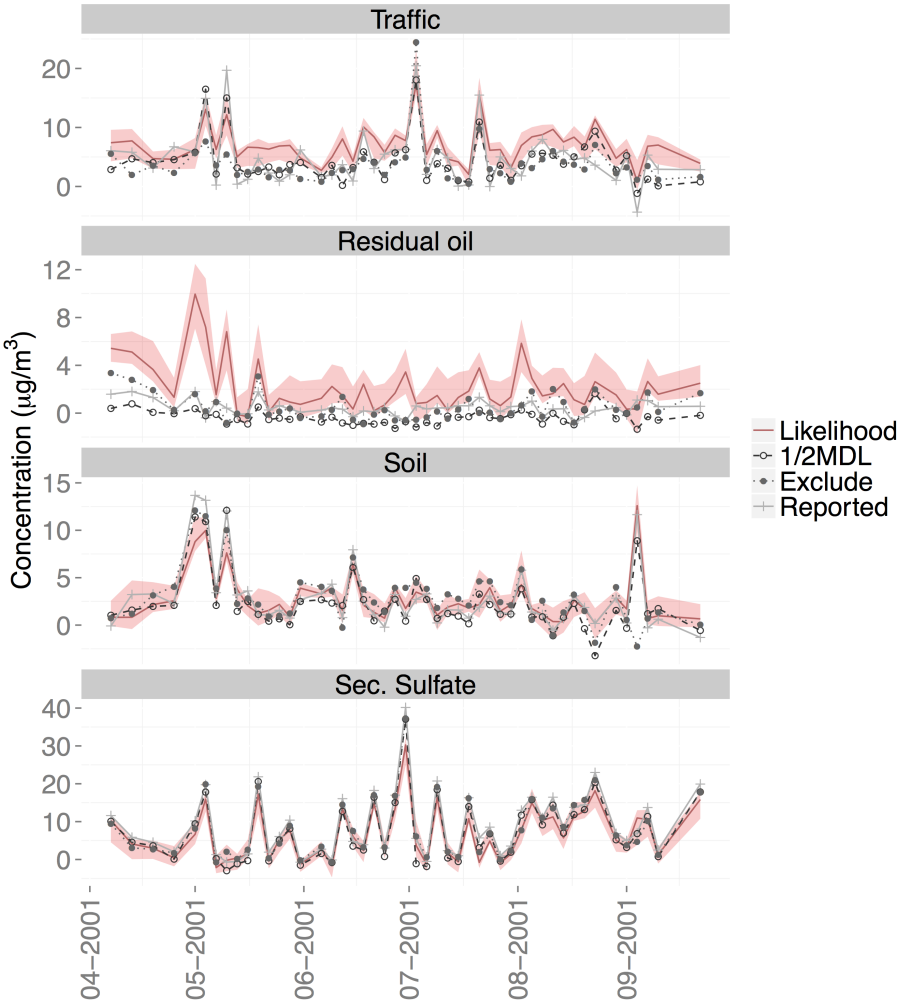

Figure 5.2: New York City time series of sources estimated using APCA from April-September 2001. Time series were estimated using different censoring adjustment methods: $\frac{1}{2}\times$MDL, Exclude, and Reported values below the MDL. Also shown is the interquartile range of estimated time series using the likelihood approach for multiple draws from the truncated lognormal distribution.

Table 5.17: Mean concentrations (standard deviations) in $\mu g/m^3$ for four sources in New York City estimated using PMF including soil, secondary sulfate (sec. SO_4^{-2}), traffic, and residual oil/incineration. Results using four different methods for adjusting censored data are shown: Reported data, Likelihood, $\frac{1}{2}\times$MDL, Exclude/downweight.

	Soil	Sec. SO_4^{-2}	Traffic	Res. oil/incineration
Reported	1.05 (1.07)	5.47 (5.38)	2.61 (1.7)	1.61 (1.62)
Likelihood	0.93 (0.82)	5.39 (5.31)	2.24 (1.44)	1.59 (1.22)
1/2 MDL	1.35 (1.41)	5.15 (5.11)	2.61 (2.62)	0.91 (1.09)
Exclude	0.94 (1.02)	4.29 (4.78)	2.72 (2.76)	0.29 (0.36)

5.6 Discussion

We have provided the first comprehensive examination of how different censoring adjustment methods impact estimation of $PM_{2.5}$ sources. Generally sources must be estimated from available $PM_{2.5}$ constituent concentrations, which frequently are censored below MDLs. Because common source apportionment methods cannot handle missing data, guidance on how to adjust censored constituent data is critical for $PM_{2.5}$ source estimation. While many previous studies have determined the best methods for adjusting censored data when estimating summary statistics or performing traditional factor analysis or PCA (Helsel, 2005b; Farnham et al., 2002; Aruga, 1997), no studies have comprehensively examined how censored data impacts source apportionment. Most source apportionment studies do not use likelihood-based approaches to impute censored data. We demonstrated that while a likelihood-based imputation approach frequently leads to better source apportionment results, substitution methods are also appropriate when few data are censored.

Maximum likelihood methods for censored data yield results that differ from results using standard substitution methods (Chen et al., 2013; Ganser and Hewett,

2010), and generally likelihood-based methods are preferred to substitution methods (Helsel, 2010). However, the best censoring adjustment method to estimate the mean of the data may not be the best censoring adjustment method to estimate other quantities from the data (Helsel, 2010; Ganser and Hewett, 2010). We developed a likelihood-based approach for imputing censored data that allows us to examine the uncertainty in source estimation driven by censored data (e.g. Figure 5.2), which has not been previously applied in source apportionment. When many data are censored, our likelihood-based imputation method improves source estimation under APCA relative to other commonly-applied censoring adjustment methods. For PMF, we found that our likelihood-based approach for imputing censored data does not lead to source estimates that are much improved compared to using commonly applied censoring adjustment methods. Future work could develop a novel likelihood-based approach for PMF that inflates censored observation uncertainties using multiple imputation variability. Under both APCA and PMF, when only a few constituents are censored, substituting concentrations below the MDL with $\frac{1}{2}\times$MDL leads to source apportionment results similar to those obtained using a likelihood-based method.

The exclude and/or downweight methods did not perform consistently better than other censoring adjustment methods in our simulation study. In our simulation study using PMF, we assumed all data had equal uncertainties to minimize differences from APCA. Paatero and Hopke (2003) suggested that if constituents with small concentrations have small signal-to-noise ratios, dropping or downweighting constituents may lead to less bias in PCA. If we generated data so that concentrations close to the MDL had larger uncertainties, downweighting constituents may have led to better source apportionment results. Previous studies have selected uncertainties as a function of the MDL and the analytical uncertainty (Song et al., 2001; Larson et al.,

2004), but it is not clear how best to generate analytical uncertainties for a simulation study. Additionally, we randomly selected constituents to be censored, while in practice the constituents with many censored concentrations may not be necessary for source estimation and excluding them may, in fact, decrease bias.

We found that as censoring increased, the estimated source means and variances obtained from the censored data were farther from those obtained from the uncensored data in a simulation study. The estimated mean bias and variance ratios due to censoring were likely underestimated because we could only compute these quantities for sources correctly identified under censoring. For example if dust and diesel in the 3-source scenario were classified as coal and vehicle under censoring, we were unable to determine which source was dust. The mean bias and variance ratios were conditional on our ability to match the sources estimated from the censored data with the true sources. However, we were not able to completely eliminate misclassification in the computation of the mean bias and variance ratios. In the 5-source scenario under APCA, the variance of diesel was overestimated under censoring and the variance of vehicle was underestimated under censoring (Table 5.7). These extreme values likely occurred because vehicle and diesel were switched in the classification step and the variance used to generate the diesel source was much smaller than the variance used for the vehicle source (Table 5.2). This example reflects possible reporting errors in practice where the source mean and variance could be incorrectly estimated because the source was incorrectly misclassified. This also reflects some difficulty in comparing simulation study results between APCA and PMF. The results from the simulation study indicate that PMF misclassifies sources more frequently than APCA, but the mean bias and variance ratios are frequently better for PMF than APCA. It is likely that PMF misclassifies sources under censoring, leading to better

estimates of the mean and variance conditional on correct classification.

We did not directly compare APCA and PMF, though other studies have found that source apportionment results were similar across methods (Ito et al., 2004; Hopke et al., 2006; Rizzo and Scheff, 2007; Lingwall et al., 2008). Our work adds to the body of research comparing APCA and PMF by demonstrating that censored data should be treated differently depending on the source apportionment model. Instead of using APCA or PMF, we could directly modify the source apportionment model to handle censored data. Tobit factor analysis can be applied to censored data, but does not yield non-negative results and is not frequently applied to environmental data (Muthén, 1989; Kamakura and Wedel, 2001). Additionally, a fully Bayesian factor analysis model (Lingwall et al., 2008; Nikolov et al., 2011) could be fitted to the observed and censored data that yields non-negative results.

In most time series studies of the short-term health effects of $PM_{2.5}$ sources (Laden et al., 2000; Mar et al., 2006), source apportionment is first used to estimate $PM_{2.5}$ sources. Then the associations between estimated source concentrations and daily adverse health counts, such as mortality, are estimated using log-linear regression. Measurement error in source concentration estimates can lead to biased health effect estimates in this two-stage approach (Zidek et al., 1996; Fung and Krewski, 1999). In Section 5.3, we demonstrated that the method used to adjust censored data can lead to errors in source estimation. Our simulation study results indicated that source concentration variances estimated from source apportionment differ between censored and observed data, particularly as the amount of censored data increases (Tables 5.6, 5.7, 5.10, 5.11). By using our likelihood-based approach to estimate sources prior to estimating source-specific health effects, we can propagate the uncertainty in source estimation driven by censoring to estimate the precision of our

health effect estimate.

We have demonstrated how different censoring adjustment methods impact source apportionment results. When many data were censored, a likelihood-based approach to impute censored data improved source estimation. When few data were censored, substituting chemical constituent concentrations with $\frac{1}{2}\times$MDL led to good estimates of source means and variances. In general, excluding or downweighting constituents with many censored concentrations did not improve source estimation. We estimated $PM_{2.5}$ sources in New York City and found estimated source means and variances differed by censoring adjustment method for APCA, but were generally similar across methods for PMF. Estimation of $PM_{2.5}$ emitted from different sources can be impacted by the method chosen to impute or remove censored $PM_{2.5}$ constituent concentrations. Therefore careful selection of censoring adjustment methods in source apportionment is necessary.

5.7 Supplementary material

The US EPA SPECIATE database (version 4.2), contains over 2,500 chemical constituent profiles of $PM_{2.5}$ sources from across the US. Each source profile consists of percent contributions from 53 chemical constituents, including the 23 constituents used in our simulation study as well as the 30 constituents used in the analysis of $PM_{2.5}$ sources in New York City. We used the SPECIATE database in our simulation study to select prototypical profiles and to identify source types. Before using the SPECIATE database in our study, we excluded all profiles that were "composite" or "average" profiles, limiting the database to estimated source profiles obtained from different locations across the US.

In SPECIATE, the profiles are each named for their source, but the names are

not consistently applied to all profiles. Therefore, we created broader categories to represent broader source definitions for our simulation study. We identified key words associated with seven source categories prominent in the dataset: wood burning (wood), road dust (dust), diesel exhaust (diesel), motor vehicles (vehicle), coal combustion (coal), metals production (metal), and oil combustion (oil). Then, we examined the source names that included each of the key words and removed profiles that did not fit in the category. We examined the final list of profiles for each source and inspected the names to ensure that they were properly classified. The key terms used to identify each source, as well as those key terms excluded, are given in Table 5.18.

Table 5.18: Sources chosen from SPECIATE including key words and words excluded.

source name	key words	removed
wood burning	wood, burning	sander dust
diesel exhaust	diesel	
road dust	soil, dust, crustal, sand, road dust particulate, earth, dirt, paved road	coal dust
motor vehicles	vehicle, gasoline, gas combustion	leaded, diesel
coal combustion	coal	
metals production	metal, steel, copper, lead	leaded
oil combustion	oil, petroleum refinery	

The seven source categories had respectively: wood (196 profiles), diesel (212 profiles), dust (708 profiles), vehicle (318 profiles), coal (113 profiles), metal (111 profiles), oil (86 profiles). We limited each profile to the 23 constituents used in the simulation study. Then, we rescaled the profiles so that the entries for the 23 constituents represented the percent contributions to each source.

Table 5.19: Constituents that contribute substantially to each of the 5 sources from the simulation study.

source type	constituents
wood	OC, EC, potassium, sulfate
diesel	OC, EC, sulfate, silicon
dust	titanium
vehicle	OC, EC, potassium
coal	titanium

5.7.1 Choosing profiles from SPECIATE for the simulation study

We chose prototypical profiles from our cleaned SPECIATE dataset to generate constituent concentration data for our simulation study. We identified several constituents that contribute substantially to each of the five sources used in the simulation study as shown in Table 5.19. Then, we randomly selected profiles from each source type that had large contributions from those constituents. While the different source types shared constituents that contributed most to the source, the sources were inherently different as noted by the name classification in SPECIATE and the full chemical profile.

5.7.2 Classifying profiles using SPECIATE

We used a k-nearest neighbors approach to identify estimated sources that first finds the k profiles in SPECIATE closest in Euclidean distance to the estimated profile. Then, k-nearest neighbors classifies each estimated source profile using the majority of those k closest profiles. We did not want two source profiles from the same dataset to be classified as the same source, since this would not reflect observed $PM_{2.5}$ sources. Therefore we reclassified duplicated source profiles, those that had the smaller number of k closest profiles corresponding to the source, based on the remaining sources. For example, if two source profiles were both classified as wood,

the source profile with the smaller number of k closest wood profiles in SPECIATE would be reclassified using non-wood source profiles.

In order to select k for each source scenario, we applied source apportionment to each simulated, uncensored dataset, $\mathbf{X}$. Then, we classified the estimated source profiles using k-nearest neighbors for k$\in \{1, 10, 20, 50, 100\}$. We chose k such that the estimated profiles from source apportionment for the uncensored data were correctly classified. We then determined the average number of misclassified sources by applying k-nearest neighbors to the source profiles estimated under censoring for each simulated dataset. We chose k to maximize correct source classification under no censoring: k$=1$ for three sources and k$=20$ for five sources. With 5 sources, it was difficult to classify profiles as 5 separate sources with no duplicates using k$=1$ since our cleaned SPECIATE database had only 7 total source categories.

Chapter 6

A method to identify regional particulate matter sources and their health effects

Determining whether different sources of particulate matter (PM) air pollution vary in toxicity is critical for the study of PM pollution. Sources of PM are not directly measured and frequently must be inferred from PM chemical constituent concentrations observed at ambient monitors. To estimate regional associations between PM sources and adverse health outcomes, it is necessary to pool estimated health effects across monitors. Pooling estimated health effects of PM sources is challenging because PM sources are frequently estimated separately for each ambient monitor and the sources that generate PM vary between communities. Currently, ad hoc approaches are applied to pool estimated health effects of PM sources across monitors, but these methods become infeasible for large, regional studies. We developed a novel approach for identifying major PM sources shared across multiple monitors that guides pooling source information, such as estimated health effects, across monitors. First, our method estimates the chemical composition of PM sources at individual monitors using a principal component analysis (PCA) approach. Then, the method extracts

major PM sources using a second-level PCA applied to the chemical composition of PM sources from all monitors. The resulting database of major PM sources is used to guide pooling source information across multiple monitors. Using data from 2000-2005 for 24 communities in the northeastern US, we applied our method to estimate the first regional associations between short-term exposure to major PM sources and mortality.

6.1 Introduction

Exposure to particulate matter (PM) air pollution has been associated with increased risk of mortality and morbidity (Dominici et al., 2006; Mar et al., 2000; Pope et al., 2002; Ostro et al., 2006; Zanobetti and Schwartz, 2009) and PM less than 2.5 μm in aerodynamic diameter ($PM_{2.5}$) is likely more toxic than other size fractions of PM (Zanobetti and Schwartz, 2009; Environmental Protection Agency, 2009). To estimate regional and national health effects of $PM_{2.5}$, community-level health effects are frequently pooled across multiple communities (Zanobetti and Schwartz, 2009; Peng et al., 2009; Ostro et al., 2006). Pooling health effects of $PM_{2.5}$ across communities increases precision in estimated health effects (Samet et al., 2000a,b), which is important because some effects associated with PM exposure are small in magnitude and can be difficult to estimate with data from only one community. A critical gap in the study of $PM_{2.5}$ is whether $PM_{2.5}$ from different sources varies in toxicity. In general, sources of $PM_{2.5}$ are not directly measured and are inferred from the chemical composition of $PM_{2.5}$ using source apportionment models. Source apportionment models are frequently applied to data from one monitor and ad hoc approaches are used to pool information across multiple monitors. These ad hoc approaches for pooling information are infeasible for large studies with many ambient monitors because pairwise comparisons of source information from individual monitors are usually necessary. However, precise estimation of health effects associated with exposure to $PM_{2.5}$ sources likely requires pooling estimated community-level health effects across multiple communities.

For data from one monitor, source apportionment models use the observed $PM_{2.5}$

constituent matrix $\mathbf{X}_{[T\times P]}$ to estimate two nonnegative matrices: the source concentration matrix $\mathbf{F}_{[T\times L]}$, which represents the concentration of each unobserved source l on day t and the source profile matrix $\mathbf{\Lambda}_{[L\times P]}$, which describes the relative contribution of each chemical constituent p to each source l. The source profile matrix characterizes sources and is used to link estimated sources to known sources of pollution at that monitor. The time series from the source concentration matrix $\mathbf{F}_{[T\times L]}$ can be used in regression models to estimate community-level associations between sources and adverse health outcomes. Common source apportionment methods include Positive Matrix Factorization (PMF) (Paatero and Tapper, 1994; Norris et al., 2008), Unmix (Henry, 1997; Norris et al., 2007), Absolute Principal Component Analysis (APCA) (Thurston and Spengler, 1985) and Bayesian models (Nikolov et al., 2011; Lingwall et al., 2008). Source apportionment models differ from traditional PCA and factor analysis because they aim to estimate non-negative $\mathbf{F}$ and $\mathbf{\Lambda}$.

Because source apportionment results cannot be easily pooled across communities, most source apportionment studies estimate $PM_{2.5}$ sources for a single community using data from one ambient monitor (Ito et al., 2006; Laden et al., 2000; Sarnat et al., 2008). However, if a source is present at multiple neighboring monitors, pooling information will lead to better estimates of the source by decreasing the impact of outlying monitors. Estimating health effects associated with $PM_{2.5}$ sources across larger regions will lead to more precise estimates. Precise national-level health effect estimates are necessary to consider possible changes to $PM_{2.5}$ regulation (Environmental Protection Agency, 2009). Additionally, if we are interested in community-specific health effects, reporting empirical Bayes estimates that pool information across monitors will lead to better community-specific estimates (Carlin and Louis, 2009).

Combining source information across ambient monitors is challenging because the presence of $PM_{2.5}$ sources and the chemical composition of $PM_{2.5}$ sources may vary between communities. As an example, Larson et al. (2004) identified a vegetative burning source in Seattle, WA that is not present in New York City, NY (Ito et al., 2004). The chemical composition of $PM_{2.5}$ sources may depend on factors that vary between communities, such as traffic-related $PM_{2.5}$ varying in chemical composition with the proportion of diesel engines. Another challenge in combining source information is that source apportionment models applied to two ambient monitors could yield sources in a different order, e.g. the columns of $\mathbf{F}_{[T\times L]}$ could be switched. For two collocated monitors, source apportionment might identify sources for monitor 1 ordered as (coal combustion, traffic, road dust), while results for monitor 2 might be ordered as (traffic, coal combustion, road dust). Ad hoc approaches are generally used to match source apportionment results between monitors, guided by information such as the sources' chemical compositions and the temporal correlations between source concentrations (Ito et al., 2004; Bell et al., 2013). Because these ad hoc approaches require pairwise comparisons to be made between all monitors, these approaches become infeasible as the number of monitors increases.

We developed a novel approach for identifying sources of $PM_{2.5}$ that are SHared Across a REgion, a method we call SHARE. This paper demonstrates the utility of SHARE for pooling source apportionment results across multiple ambient monitors. SHARE first identifies major $PM_{2.5}$ sources which we define as those sources that are present at multiple monitors and explain much of the variability in the data. Then using this database of major $PM_{2.5}$ sources, SHARE determines which major sources are present at each individual monitor. In a simulation study, we compared pooling source apportionment results using SHARE to a method that assumes all $PM_{2.5}$

sources are the same across multiple ambient monitors. Using data from 2000-2005 for 24 urban communities in the northeastern US, we used SHARE to estimate major $PM_{2.5}$ sources and to estimate the first regional associations between daily non-accidental, all-cause mortality and short-term exposure to $PM_{2.5}$ sources.

6.2 Data

The US Environmental Protection Agency's Chemical Speciation Network (EPA CSN) is a national monitoring network of approximately 250 ambient $PM_{2.5}$ speciation monitors that measure daily concentrations for over 50 $PM_{2.5}$ chemical constituents. We restricted our analysis to 24 chemical constituents of $PM_{2.5}$ (Table 6.1) that contribute to previously identified $PM_{2.5}$ sources (Ito et al., 2004; Maykut et al., 2003; Rizzo and Scheff, 2007). These constituents include major ions (e.g. sulfate and nitrate), metals (e.g. zinc and vanadium), and carbon-containing constituents (elemental carbon (EC) and organic carbon (OC)). For the six year period from 2000-2005, we created a dataset of 41 speciation monitors in northeastern US communities, defined as a county or set of counties containing an urban area. Each of these 41 monitors had more than 50 daily observations for all 24 $PM_{2.5}$ chemical constituents. This region contains communities with potentially different sources of $PM_{2.5}$, including coastal, industrial, and heavily populated communities. The locations of the 41 $PM_{2.5}$ speciation monitors used in this study are shown in Figure 6.1.

To estimate associations between $PM_{2.5}$ sources and mortality, we used daily counts of all-cause, non-accidental mortality from the National Center for Health Statistics. Our mortality dataset consisted of 24 urban communities with at least one $PM_{2.5}$ speciation monitor. All 41 speciation monitors from our restricted EPA CSN

Table 6.1: The 24 $PM_{2.5}$ chemical constituents used to estimate $PM_{2.5}$ sources in this analysis.

Aluminum	Ammonium	Arsenic	Bromine	Calcium
Chlorine	Copper	Elemental Carbon (EC)	Iron	Potassium
Manganese	Sodium ion	Nickel	Nitrate	Organic Carbon (OC)
Phosphorus	Lead	Selenium	Silicon	Sulfate
Strontium	Titanium	Vanadium	Zinc	

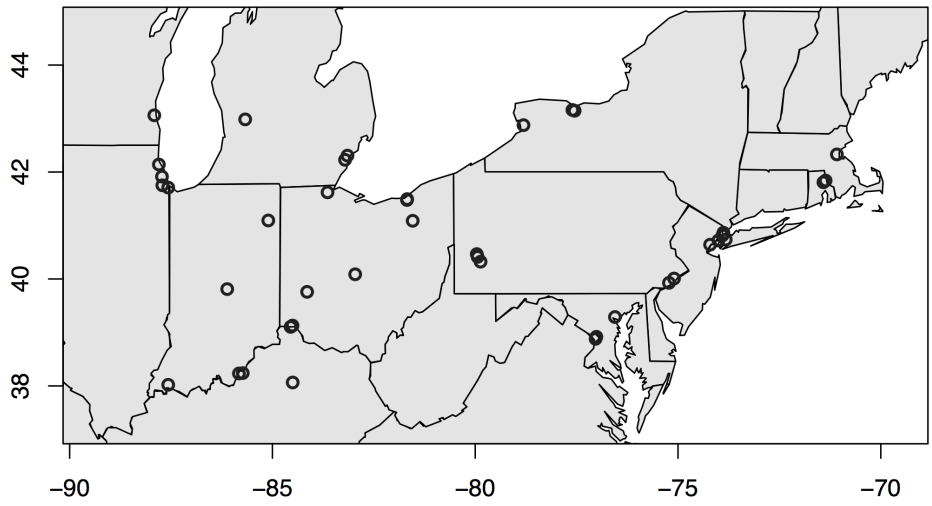

Figure 6.1: Map of 41 $PM_{2.5}$ chemical constituent monitors from US EPA chemical speciation network used in this analysis.

dataset fall within one of these 24 communities. We also obtained daily temperature for each community from the National Oceanic and Atmospheric Administration (EarthInfo Inc., 2006).

6.3 Methods

6.3.1 SHared Across a REgion (SHARE) method

We developed the SHared Across a REgion (SHARE) method to estimate major $PM_{2.5}$ sources and to determine which monitors observe each major source. Many studies of $PM_{2.5}$ sources at individual ambient monitors use principal component analysis (PCA) to determine the chemical makeup of each source of $PM_{2.5}$. Consider each principal component (PC) as a "source signature," which identifies the chemical constituents that are present in that source. Taking together the collection of source signatures obtained separately from multiple monitors, there will be some duplicated source signatures if $PM_{2.5}$ sources are shared between monitors. By applying a second-level PCA to the concatenated source signatures from all monitors as in Population Value Decomposition (Crainiceanu et al., 2011), we can determine the major $PM_{2.5}$ sources that explain most of the variability across all source signatures. This second-level PCA yields a database of major $PM_{2.5}$ sources, which can be matched with source signatures from each individual monitor to determine whether major sources are present at each monitor.

Step (1) of SHARE estimates the source signatures at each monitor by applying a varimax rotated PCA to data from each monitor. For monitor i, let $\mathbf{X}$ be the $[T_i \times P]$ matrix of $PM_{2.5}$ chemical constituent concentrations, with x_{tp} the concentration for constituent p on day t. Let $z_{tp} = \frac{x_{tp} - \bar{x}_p}{s_p}$ where $\bar{x}_p$ is the sample mean concentration for constituent p and s_p is the corresponding sample standard deviation. Then, the

vectors of length T_i that make up the $[T_i \times P]$ matrix $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_P)$ for each monitor are all mean zero with unit variance. Let $\mathbf{Z}_i$ represent the data for monitor i. We obtained the source signatures for each monitor i in the following way:

i Apply PCA to $\mathbf{Z}_i$ and select the number of sources L_i, the number of eigenvalues of $\mathbf{Z}_i^T \mathbf{Z}_i$ greater than 1.

ii Obtain $\tilde{\mathbf{V}}_i$, the matrix of the first L_i PCs, where the norm of each column is equal to the corresponding squared eigenvalue.

iii Find $\mathbf{V}_i = \tilde{\mathbf{V}}_i \mathbf{R}_i$, where $\mathbf{R}_i$ is the $L_i \times L_i$ varimax rotation matrix that satisfies

$\mathbf{R}_i = \arg\max_{\mathbf{R}} \sum_{p=1}^{P} \sum_{l=1}^{L_i} (\tilde{\mathbf{V}}_i \mathbf{R})_{pl}^4 - \frac{1}{P} \sum_{l=1}^{L_i} (\sum_{p=1}^{P} (\tilde{\mathbf{V}}_i \mathbf{R})_{pl}^2)^2$ (Harris and Kaiser, 1964) and where P is the number of constituents

The varimax rotation maximizes the sample variance for each PC, which creates more interpretable source signatures by pushing the absolute values of the PC loadings closer to 0, representing constituents that are not present in the source, or 1, representing constituents present in the source. For each monitor i, the $[P \times L_i]$ matrix $\mathbf{V}_i$ represents the source signatures.

Steps (2) and (3) in SHARE create a database of major $PM_{2.5}$ sources. In Step (2) we concatenated all source signatures $\mathbf{V}_i$ across N monitors $\mathbb{V} = [\mathbf{V}_1, \mathbf{V}_2, ..., \mathbf{V}_N]^T$. This matrix $\mathbb{V}_{[\sum_{i=1}^{N} L_i \times P]}$ consists of all source signatures across all monitors. If monitor 1 and monitor 2 share a $PM_{2.5}$ source, then there will be at least one similar column in both $\mathbf{V}_1$ and $\mathbf{V}_2$, and $\mathbb{V}$ will be nearly rank deficient. Then in Step (3), we reduced our matrix $\mathbb{V}$ down to unique, major $PM_{2.5}$ sources, which are defined as those sources that explain most of the variability in $\mathbb{V}$ and will be sources that are present at multiple monitors. To accomplish this, we applied varimax-rotated PCA

to the centered and scaled $\mathbb{V}$ and retained the first L components. We choose L based on the number of eigenvalues of $\mathbb{V}$ that are greater than 1, which is common in PCA and corresponds to keeping PCs that explain more variability than one of the chemical constituents alone (Guttman, 1954; Ito et al., 2004). In practice, we found PCs with eigenvalues less than 1 were noisy and did not resemble known $PM_{2.5}$ sources. Let $\mathbb{U}_{[P\times L]}$ be the matrix of the first L rotated PCs of $\mathbb{V}$ representing the major $PM_{2.5}$ source signatures. Thus far, SHARE has estimated source signatures for each monitor, $\mathbf{V}_i$, and the source signatures for major $PM_{2.5}$ sources shared across multiple monitors, $\mathbb{U}$.

Step (4) of SHARE identifies which major $PM_{2.5}$ sources are present at each monitor. In factor analysis, it is common to check the similarity between two factors using the congruence correlation, or the cosine of the angle between two factors (Harman, 1976). To match major $PM_{2.5}$ sources to source signatures at monitor i, we first found the angles between all pairs of sources and created a $[L_i \times L]$ matrix of angles between the L_i sources from monitor i and the L major sources. The angle between source l at monitor i and major source l' is $arccos\left(\frac{\mathbf{V}_i(l)^T\mathbb{U}(l')}{\|\mathbf{V}_i(l)\|\|\mathbb{U}(l')\|}\right)$, where $\mathbf{V}_i(l)$ is the source signature for source l at monitor i and $\mathbb{U}(l')$ is the source signature for major source l'. Small values in the angle matrix correspond to sources at monitor i that are similar to major sources in $\mathbb{U}$. Because source signatures at each monitor are considered to be unique, we did not want two source signatures from $\mathbf{V}_i$ to be matched to the same major $PM_{2.5}$ source in $\mathbb{U}$. Therefore, we used the Hungarian method (Papadimitriou and Steiglitz, 1998; Kuhn, 1955) to determine the best matches between source signatures from $\mathbf{V}_i$ and major sources in $\mathbb{U}$ without duplicating matches. The Hungarian method finds the optimal assignment between sources in $\mathbf{V}_i$ and $\mathbb{U}$ that minimizes the sum of the corresponding elements of the angle matrix. Some sources

at monitor i will not be major $PM_{2.5}$ sources and some major sources will not be present at monitor i, so we only allowed matches for angles less than 45 degrees. This cutoff ensured matched source signatures at monitor i were closer to the major $PM_{2.5}$ source than to a vector orthogonal to the major source.

The four major steps of SHARE are depicted in Figure 6.2. Step (1) finds source signatures at each monitor by applying PCA, but at this point the major sources present at each monitor are unknown (represented by open rectangles). Step (2) concatenates source signatures across all monitors. Step (3) finds major sources by applying a second-level PCA to the source signatures. At this point because we have reduced the number of source signatures, we can determine the nature of each major $PM_{2.5}$ sources (represented by closed rectangles). Step (4) matches major sources with source signatures at each monitor by computing the angle between all sources and using the Hungarian method for optimal assignment. Note the "black" source is only present at monitor 2 and is therefore not a major $PM_{2.5}$ source. SHARE yields chemical signatures for major $PM_{2.5}$ sources shared across multiple monitors and also determines whether major sources are present at each monitor. Knowing which monitors observe a particular source would enable source information, such as estimated health effects, to be pooled across multiple monitors.

6.3.2 Estimating source concentrations

For each ambient speciation monitor, we estimated daily concentrations for $PM_{2.5}$ sources using APCA, a commonly applied source apportionment method (Hopke et al., 2006; Ito et al., 2004) that can be easily implemented in common statistical packages. APCA first finds absolute principal component scores $\mathbf{A}$ by rotating and rescaling results from PCA. Let $\mathbf{A}_{[T_i \times L_i]} = (\mathbf{X}\mathbf{S}^{-1})[\mathrm{Cor}(\mathbf{X})^{-1}\mathbf{V}]$, where

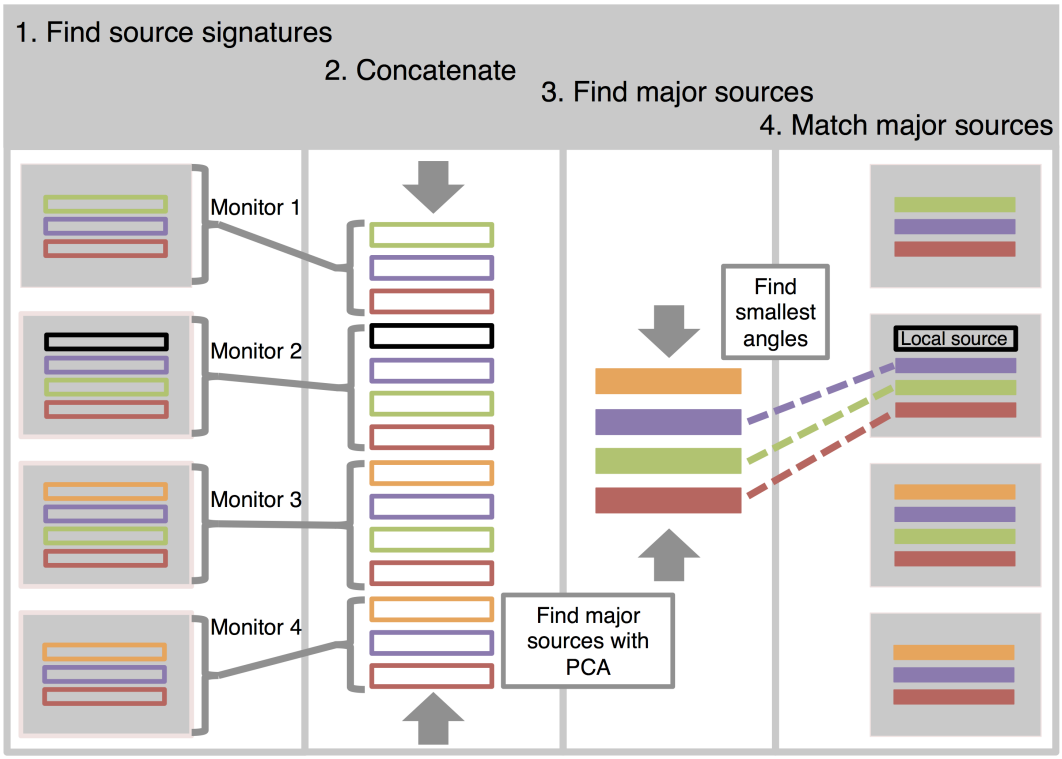

Figure 6.2: Conceptual picture illustrating the four major steps of SHARE.

$\mathbf{S} = \text{diag}(s_1, s_2, ..., s_P)$ is a diagonal matrix of the sample standard deviations of the columns of the constituent data $\mathbf{X}$ and where $\mathbf{V}$ is the source signature for one monitor computed in steps i-iii (Section 6.3.1). The absolute principal component scores $\mathbf{A}$ are the scaled but uncentered data rotated into the factor space. In the next step, daily total mass $PM_{2.5}$ is regressed on $\mathbf{A}$

$$PM_t = \eta_0 + \mathbf{a}_t^T \boldsymbol{\eta} + \varepsilon_t \tag{6.1}$$

The daily concentrations for each source l are estimated as $f_{tl} = a_{tl} \times \hat{\eta}_l$. APCA yields estimated source concentrations for each individual monitor and ad hoc approaches are necessary to pool APCA results across multiple monitors. By combining APCA with SHARE, as described in detail in the next section, we can estimate regional health effects for sources estimated using APCA.

Thurston et al. (2011) proposed a mixed modeling approach to expand APCA to multiple monitors, which we refer to as multiple APCA (mAPCA). This mAPCA method concatenates all $PM_{2.5}$ constituent data across monitors to find the absolute principal component scores, and therefore assumes $PM_{2.5}$ sources are the same across monitors. Let $\mathbf{X}'$ be the $[\sum_i T_i \times P]$ matrix of concatenated data from all monitors i. First, mAPCA computes the source signatures $\mathbf{V}'_{[P \times L']}$ from steps i-iii in Section 6.3.1 using $\mathbf{X}'$, where the number of sources L' may be different from L. Then, the mAPCA absolute principal component scores are $\mathbf{A}'_{[\sum_i T_i \times L']} = (\mathbf{X}'\mathbf{S}'^{-1})[\text{Cor}(\mathbf{X}')^{-1}\mathbf{V}']$, where $\mathbf{S}' = \text{diag}(s'_1, s'_2, ..., s'_P)$ and s'_p is the standard deviation of constituent p for the concatenated dataset $\mathbf{X}'$. To estimate source concentrations, mAPCA fits a mixed model with a random intercept for monitor and regresses the concatenated total mass $PM_{2.5}$ from all monitors on $\mathbf{A}'$,

$$PM_{tm} = \xi_0 + b_m + \mathbf{a}'^T_{tm} \boldsymbol{\xi} + \varepsilon_{tm} \tag{6.2}$$

where b_m is the random intercept, t is the observation day, and m is an individual monitor. Then the source concentration corresponding to source l for monitor m on day t is $f_{tm,l} = a'_{tm,l} \times \hat{\xi}_l$. Because the sources for each vector $\mathbf{a}'_{tm}$ are the same regardless of the monitor m, mAPCA assumes sources are the same across monitors. We implemented APCA and mAPCA using R version 3.0.2 (R Core Team, 2012).

6.3.3 Estimating associations between $PM_{2.5}$ sources and mortality

We estimated community-level associations between short-term exposure to $PM_{2.5}$ sources and mortality using a two-stage approach that has been previously applied to study associations between mortality and $PM_{2.5}$ sources in Phoenix, AZ and Washington, DC (Mar et al., 2006; Ito et al., 2006). After estimating daily community-level $PM_{2.5}$ source concentrations $\mathbf{f}_t$ using either APCA or mAPCA, we estimated associations between daily mortality and each $PM_{2.5}$ source l using a log-linear time series model for each community,

$$\begin{aligned} y_t &\sim Poisson(\mu_t) \\ \log(\mu_t) &= \beta_0 + f_{tl}\beta_l + \mathbf{w}_t^T \boldsymbol{\nu} \end{aligned} \tag{6.3}$$

For a particular community on day t, y_t is the number of deaths and $\mathbf{w}_t$ is the vector of covariates. For each community, we estimated associations between $PM_{2.5}$ sources and mortality using the log relative risks $\hat{\boldsymbol{\beta}}$ and their corresponding standard errors.

We then estimated regional associations between $PM_{2.5}$ sources and mortality. To estimate regional mortality effects, we fitted a two-level Bayesian hierarchical model separately for each estimated $PM_{2.5}$ source as in previous epidemiologic studies of PM (Peng et al., 2009; Krall et al., 2013). The Bayesian hierarchical model assumes the estimated log relative risks for each source l and community c, $\hat{\beta}_{lc}$, are normally

distributed and centered around the true log relative risk β_{lc},

$$\hat{\beta}_{lc} \sim N(\beta_{lc}, \hat{\sigma}_{lc}^2)$$

where $\hat{\sigma}_{lc}$ is the estimated standard error of $\hat{\beta}_{lc}$ from the mortality risk regression model (Equation 6.3) and is assumed to be known. In the second level, the community-specific log relative risks β_{lc} follow a normal distribution with mean θ_l, the regional log relative risk.

$$\beta_{lc} \sim N(\theta_l, \phi_l^2) \tag{6.4}$$

We used the TLNise software (Everson and Morris, 2000) implemented in R to fit the hierarchical model and to obtain regional mortality risk estimates, $\hat{\theta}_l$, and corresponding 95% posterior intervals. TLNise takes as inputs the vector of estimated community-level associations $\hat{\boldsymbol{\beta}}_l$ and corresponding estimated variances $\hat{\boldsymbol{\sigma}}_l^2$ for each source l.

We obtained the vector of community-level associations for each source differently for APCA and mAPCA. APCA is applied to data separately for each monitor and no method exists for combining APCA results across multiple monitors. Therefore, we applied SHARE to determine which sources estimated by APCA can be pooled across multiple monitors. For example, suppose SHARE identified source l at monitors 1, 2, and 5, but not at monitors 3 and 4. Then, the vector of community-level associations for source l for APCA with SHARE would be $\hat{\boldsymbol{\beta}}_l = \{\hat{\beta}_l^1, \hat{\beta}_l^2, \hat{\beta}_l^5\}$. Recall that mAPCA assumes sources are the same across monitors. To create the vector of community-level associations for the first source l from mAPCA, we selected the first source from each monitor $\hat{\boldsymbol{\beta}}_l = \{\hat{\beta}_l^1, \hat{\beta}_l^2, \hat{\beta}_l^3, \hat{\beta}_l^4, \hat{\beta}_l^5\}$.

We also used the results from the two-level Bayesian hierarchical model to estimate associations between $PM_{2.5}$ sources and mortality for each community. We estimated associations using the Bayesian posterior distribution of the community-specific associations,

$$\beta_{lc} \mid \hat{\beta}_{lc}, \hat{\theta}_l \;\; \sim \;\; N\left(H\theta_l + (1-H)\hat{\beta}_{lc},\ (1-H)\hat{\sigma}_{lc}^2\right)$$

where $\hat{\beta}_{lc}$ and $\hat{\sigma}_{lc}$ are the maximum likelihood estimate and corresponding standard error from the log-linear mortality risk regression model for source l and community c, $H = \frac{\hat{\sigma}_{lc}^2}{\hat{\sigma}_{lc}^2+\phi_l^2}$, and θ_l and ϕ_l^2 are the second-level mean and variance from the normal distribution in equation 6.4. Because both θ_l and ϕ_l^2 are unknown, we used their posterior means from the Bayesian hierarchical model to create empirical Bayes mortality risk estimates for each community and source.

6.4 Simulation study

In a simulation study, we tested the performance of SHARE. First, we determined how well SHARE identified major $PM_{2.5}$ sources across multiple monitors. Second, we compared estimated regional mortality effects between two methods for estimating sources across monitors: traditional APCA using SHARE and mAPCA, which assumes all sources are the same across monitors.

We simulated data from multiple monitors and allowed the sources present at each monitor to vary. For our simulation study, we used common sources reported in the literature including wood burning (wood), diesel exhaust (diesel), road dust (dust), motor vehicles (vehicle), and coal combustion (coal). For each monitor, we

Table 6.2: Subregions with varying sources for simulation study

Subregion	Number of sources	Sources				
I	3	wood	diesel	dust		
II	3		diesel	dust	vehicle	
III	4	wood	diesel	dust	vehicle	
IV	4		diesel	dust	vehicle	coal
V	5	wood	diesel	dust	vehicle	coal

simulated data from one of 5 subregions, or areas with different sources (Table 6.2). Each simulated dataset contained multiple monitors in the same subregion and/or multiple monitors in different subregions.

For a monitor in a given subregion, we simulated $PM_{2.5}$ constituent concentrations by taking the product of a source concentration matrix, $\mathbf{F}$, and a source profile matrix, $\mathbf{\Lambda}$,

$$\mathbf{X}_{[T\times P]} = (\mathbf{F}_{[T\times L]}\mathbf{\Lambda}_{[L\times P]}) \circ \boldsymbol{e}_{[T\times P]} \tag{6.5}$$

where $\log(e_{tp}) \overset{IID}{\sim} N(0, 0.01^2)$ represents non-negative error and $\circ$ is the Schur (element wise) product. Each row of $\mathbf{\Lambda}_{[L\times P]}$ is a source profile with rows corresponding to the sources in the respective subregion (Table 6.2). Each profile was selected from the US EPA SPECIATE database (version 4.2), which is a database of profiles for 53 chemical constituents from across the US. We limited the data to 20 constituents: the 24 constituents in Table 6.1 except chlorine, bromine, sodium ion, and ammonium because these constituents were not present in the profiles we selected. To ensure each profile represented the proportion of the source corresponding to each constituent, we rescaled profiles to sum to one. The source profiles for the simulation study are shown in Figure 6.3.

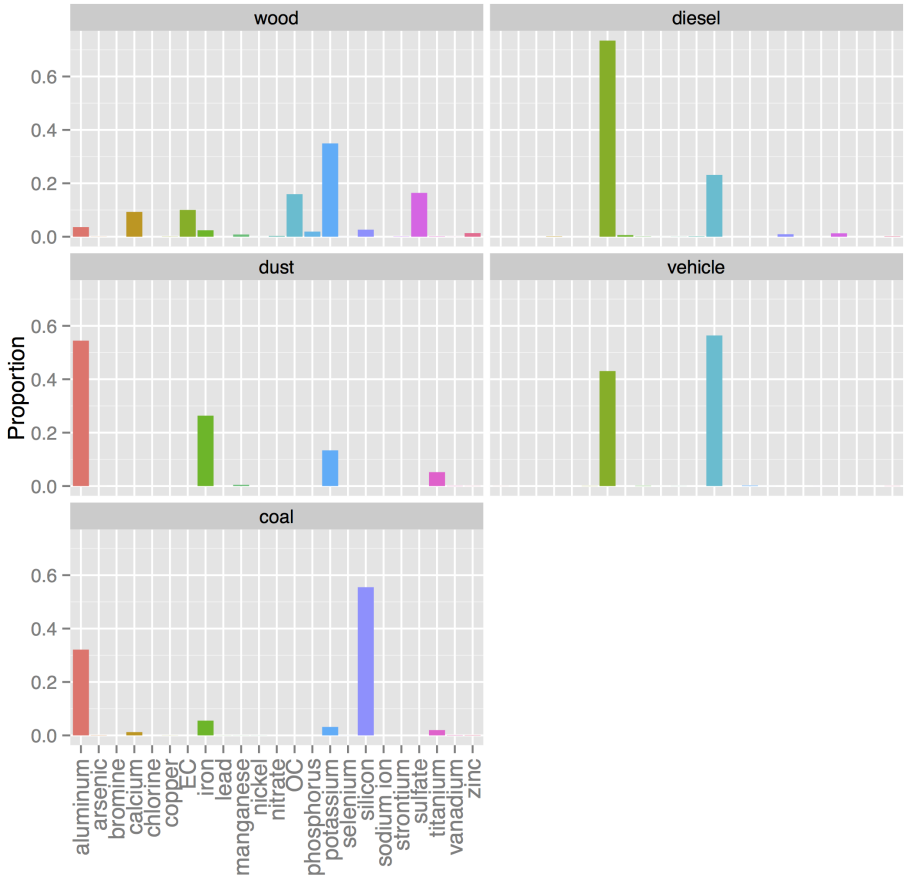

Figure 6.3: Bar plots corresponding to the profiles used in the simulation study.

Table 6.3: Means (standard deviations) of the lognormal distribution for each source in each subregion in the simulation study.

Subregion	wood	diesel	dust	vehicle	coal
I	2.56 (1.31)	5.87 (0.75)	5.14 (0.82)		
II		5.14 (0.82)	2.56 (1.31)	5.87 (0.75)	
III	2.56 (1.31)	5.87 (0.75)	5.14 (0.82)	1.73(4.33)	
IV		5.14 (0.82)	2.56 (1.31)	5.87 (0.75)	1.73(4.33)
V	2.56 (1.31)	5.87 (0.75)	5.14 (0.82)	1.73(4.33)	7.41(0.63)

For each source concentration time series $\mathbf{f}_l$, we generated $T = 1000$ independent lognormal concentrations with means and standard deviations that varied between subregions (Table 6.3). We chose these means and standard deviations to approximately reflect the distribution of sources reported in the literature (Ito et al., 2004; Lingwall et al., 2008). To select the number of sources at each monitor (L_i) and the number of major sources (L), we generally choose the number of eigenvalues of the data correlation matrix that are greater than 1 (see step i in Section 6.3.1). In the simulation study, we used the number of eigenvalues greater than 0.5, which enabled us to correctly select the number of sources. A cutoff of 1 selected too few sources because the latter sources explain less variability in the simulated data than in observed $PM_{2.5}$ constituent data. Additionally, determining how to best select the number of sources is not an aim of this analysis.

In the application of SHARE in our simulation study, when few sources are present at a monitor (e.g. 3 or fewer sources), the first source identified at the monitor matches too frequently with the first major $PM_{2.5}$ source. When PCA is applied to monitors with few sources, the first source signature (the first column of $\mathbf{V}_i$) has

many high values. These high values represent constituents that (1) contribute substantially to the first source and (2) constituents that are not generally associated with any sources at that monitor. To separate constituents in categories (1) and (2) for monitors with 3 or fewer sources, we used weights in the computation of the angle matrix between monitor-level and major sources. We selected weights of $\bar{\mathbf{x}}^{1/4}$, where $\bar{\mathbf{x}}$ is the vector of constituent means at that monitor. Because elements of $\bar{\mathbf{x}}^{1/4}$ will be small for constituents not present at the monitor, this weighting scheme down-weights source signatures for constituents that are not actually present at the monitor. We used $\bar{\mathbf{x}}^{1/4}$ instead of $\bar{\mathbf{x}}$ or $\sqrt{\bar{\mathbf{x}}}$ because the average concentration generally varies greatly between constituents and $\bar{\mathbf{x}}^{1/4}$ is less variable across constituents. When we matched monitor-level sources with major $PM_{2.5}$ sources in our simulation study using the Hungarian method, we allowed matches where the angle between sources was less than 70 degrees. Smaller cutoffs (such as 45 degrees used for observed $PM_{2.5}$ constituent data) were too stringent for the simulated data because having many monitors with few sources makes matching major $PM_{2.5}$ sources more difficult. In general, when there are many monitors with few sources, SHARE may overidentify the first source across monitors.

6.4.1 SHARE

To test whether SHARE properly identifies sources across monitors, we created a dataset of 25 monitors. We simulated data for 5 monitors in each of the 5 subregions (Table 6.2), with data from each monitor generated using equation 6.5. For each of the 25 datasets, we applied SHARE to determine which monitors had wood, diesel, dust, vehicle, and coal. In general across 100 simulated samples, SHARE correctly identified the sources at each monitor. We also tested our method with 100 monitors

(10 in each subregion) and 5 monitors (1 in each subregion), and found this did not substantially change the performance of SHARE. The results for the three datasets of different numbers of monitors ($N = 25, 100, 5$) are shown in Table 6.4. There are two possible errors when using SHARE: we identify a source at a monitor when it is truly not present (overidentified sources) or we fail to identify a true source (underidentified sources). Across 100 simulated samples, Table 6.4 shows the average number of monitors where each source was overidentified (represented by positive values) and where each source underidentified (represented by negative values). Values close to zero in the table represent cases where the source was correctly identified across monitors. As an example, under the $N = 25$ scenario, vehicle was incorrectly identified as coal for 0.12 monitors on average across 100 samples.

We also tested whether SHARE performed well in sensitivity analyses using datasets of 25 monitors as above. We first changed the number of days of data to $T_i = 5000$ for monitors in subregion III and $T_i = 200$ for monitors in subregion V. As shown in Table 6.4 (row "Days") we found results were similar to results when $T_i = 1000$ for all monitors. We also varied the number of monitors in each subregion so that subregions I-V had 5, 4, 12, 3, and 1 monitors respectively for a total of 25 monitors (Table 6.4, row "Unequal"). This allowed us to determine whether SHARE was sensitive to an uneven distribution of sources across monitors. Because coal combustion was only present at 4 monitors, coal was not a major $PM_{2.5}$ source. Therefore, SHARE could not determine whether coal combustion was present at the monitors. In this simulation, wood was incorrectly identified at 3 extra monitors (of 18 total monitors with wood) on average, but otherwise SHARE correctly identified all sources.

Table 6.4: Table of simulation study results for SHARE where each row is a different simulation. Each entry in the table corresponds to the number of monitors where the source was overidentified (positive values) or underidentified (negative values) on average across 100 samples.

Simulation	wood	diesel	dust	vehicle	coal
$N = 25$	0.00	0.00	0.00	-0.12	0.12
$N = 100$	0.00	0.00	0.00	-0.43	0.43
$N = 5$	0.02	0.00	0.00	-0.06	0.04
Days	0.01	0.00	0.00	-0.15	0.14
Unequal	3.00	0.00	0.00	0.00	-

6.4.2 Estimating mortality effects

In the second part of the simulation study, we determined whether SHARE could be used to estimate regional associations between short-term exposure to major $PM_{2.5}$ sources and mortality. We compared estimated mortality effects between mAPCA, which assumes sources are the same at each monitor, and SHARE using APCA. We simulated mortality data for each monitor as Poisson(μ_t) where $\mu_t = \exp\{5 + \sum_{l=1}^{5} \beta_l \times f_{tl}\}$. In this equation, f_{tl} are the simulated source concentrations for the monitor, β_l is the association between source l and mortality, and 5 is a background mortality rate that reflects other causes of death. We specified β_l such that the percent increase in mortality $\boldsymbol{\gamma} = 100(\exp\{10 \times \boldsymbol{\beta}\} - 1) = (3, 1, 0.75, 0.5, 1)$ corresponding to sources wood, diesel, dust, vehicle, and coal to approximate previously estimated PM-related health effects (Krall et al., 2013; Zanobetti and Schwartz, 2009; Ostro et al., 2007).

We simulated 4 datasets as part of 4 simulation scenarios, where each dataset had multiple monitors and monitors were distributed across subregions in Table 6.2. Our 4 simulation scenarios included: A (5 monitors in subregion V), B (5 monitors with 1 monitor in each subregion I-V), C (25 monitors in subregion V), D (25 monitors

with 5 monitors in each subregion I-V). In scenarios A and C, the sources are the same across all monitors and the assumption of mAPCA is met. The assumption for mAPCA is not met in scenarios B and D, where the sources present vary between monitors. Using the simulated data, we applied both APCA and mAPCA to estimate source concentrations at each monitor. We fitted log-linear regression models (equation 6.3 with no covariates) to estimate associations with mortality at each monitor. To compare both APCA and mAPCA results to those we would have obtained had we directly observed source concentrations, we regressed simulated mortality data against the known, simulated source concentrations. We obtained regional estimated mortality effects by pooling estimated mortality effects at each monitor using a two-level Bayesian hierarchical model. For APCA, we used SHARE to determine how to pool sources across monitors as described in Section 6.3.3. For mAPCA, we pooled each source successively across all monitors.

Figure 6.4 shows the percent increase in mortality and 95% posterior intervals associated with a 10-$\mu g/m^3$ increase in source concentration for APCA using SHARE (labelled as SHARE) and mAPCA. Also shown are the estimates we would have obtained if both the source locations and source concentrations were known (Known). Scenarios A and B have fewer monitors ($N = 5$) than scenarios C and D ($N = 25$) and therefore have wider 95% posterior intervals. Results from scenarios A and C, where sources are the same across monitors, are generally more precise than results for scenarios B and D, where the sources present vary between monitors. Both source estimation methods applied in this study use a PCA approach to estimate $PM_{2.5}$ sources. Since PCA aims to successively maximize the remaining variance in the data, the last few sources identified by PCA explain less variability in the data and are noisy compared with the first few sources identified by PCA. In both APCA using SHARE and

mAPCA, the source that explains the least variability in the data is vehicle exhaust. Vehicle exhaust is difficult to estimate using PCA-based approaches and therefore has wide posterior intervals and biased point estimates, particularly for mAPCA.

We performed a sensitivity analysis to determine whether the mortality simulation study results were sensitive to the magnitude of associations between mortality and short-term exposure to $PM_{2.5}$ sources. For wood, diesel, dust, vehicle and coal respectively, we compared our results to results corresponding to no associations ($\gamma_l = 0,\ l = 1...5$), all the same associations ($\gamma_l = 1,\ l = 1...5$), and both positive and null associations ($\boldsymbol{\gamma} = (1,1,0,0,0)$). We did not find the performances of APCA with SHARE and mAPCA depended on the magnitude of associations between mortality and $PM_{2.5}$ sources.

To estimate pooled community-specific mortality risk estimates, we used simulation scenario D, which had 25 monitors with varying sources across monitors (5 monitors in each subregion I-V). Scenarios A and C had the same sources across monitors and therefore we did not expect APCA with SHARE and mAPCA to perform differently. We allowed mortality risks to vary by monitor to resemble observed variation in community-specific PM health effects. For sources at a monitor, the associations with mortality were drawn from a random Uniform($\gamma_l - 0.5, \gamma_l + 0.5$), where $\boldsymbol{\gamma} = 100(\exp\{10 \times \boldsymbol{\beta}\} - 1) = (3,1,0.75,0.5,1)$ corresponding to wood, diesel, dust, vehicle, and coal. Figure 6.5 shows the empirical Bayes estimates for each monitor by source. We divided the 25 monitors by subregion to illustrate the differences in the sources present across subregions. In this scenario, wood is present at monitors in subregions I, III, and V, but not at monitors in subregions II and IV. Using APCA with SHARE, we only estimated associations for monitors where that source was identified. For monitors where wood was identified, APCA with SHARE and

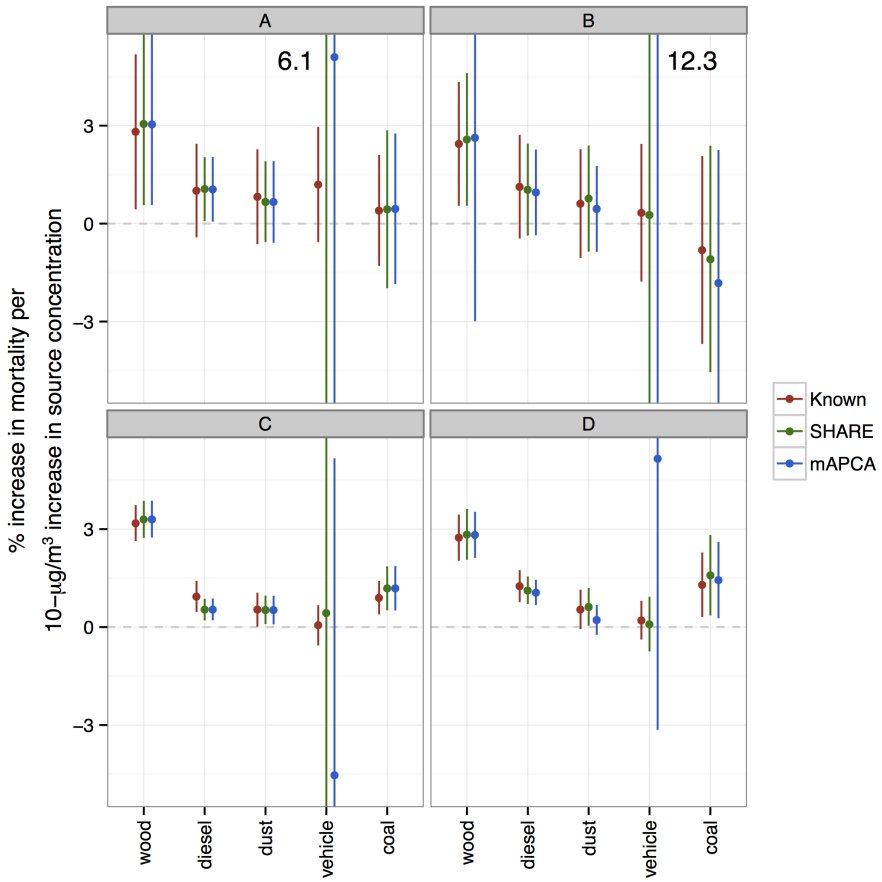

Figure 6.4: Regional percent increase in mortality (95% posterior intervals) associated with a 10 $\mu g/m^3$ increase in source concentration under 4 simulation scenarios: A (5 monitors in subregion V), B (5 monitors with 1 monitor in each subregion I-V), C (25 monitors in subregion V), D (25 monitors with 5 monitors in each subregion I-V). Each plot shows estimated effects using simulated source concentrations (Known), APCA with SHARE (SHARE), and mAPCA.

mAPCA estimated associations close to associations estimated using known source concentrations. However for monitors where wood was not present, mAPCA estimated positive, statistically significant associations between wood and mortality. Results for coal combustion were similar to results for wood burning, since coal was present in only subregions IV and V. Under mAPCA, we estimated mortality effects for coal at monitors where coal was not present, but associations were not statistically significant. Empirical Bayes estimates for vehicle had large standard errors for both APCA with SHARE and mAPCA. Additionally, mAPCA identified a statistically significant association between mortality and vehicle in subregion I, where vehicle is not present. The sources present across all monitors, diesel and dust, had similar empirical Bayes estimates for both APCA with SHARE and mAPCA.

In this simulation study, we found that SHARE can be applied to estimate major $PM_{2.5}$ sources and the monitors where these sources are present. In addition, APCA with SHARE estimated associations with mortality that are generally close to those estimated using known source concentrations.

6.5 All-cause mortality and $PM_{2.5}$ sources in the northeastern US

For 24 communities in the northeastern US, we estimated major sources of $PM_{2.5}$ and associations between short-term exposure to major $PM_{2.5}$ sources and mortality.

6.5.1 $PM_{2.5}$ sources in the northeastern US

Across 41 ambient $PM_{2.5}$ speciation monitors in our study, the number of days with complete data for all 24 constituents (Table 6.1) ranged from 56 days to 557 days with a median of 204 days. We first applied SHARE to identify the locations of major $PM_{2.5}$ sources across monitors. Table 6.5 shows the 8 major sources along with

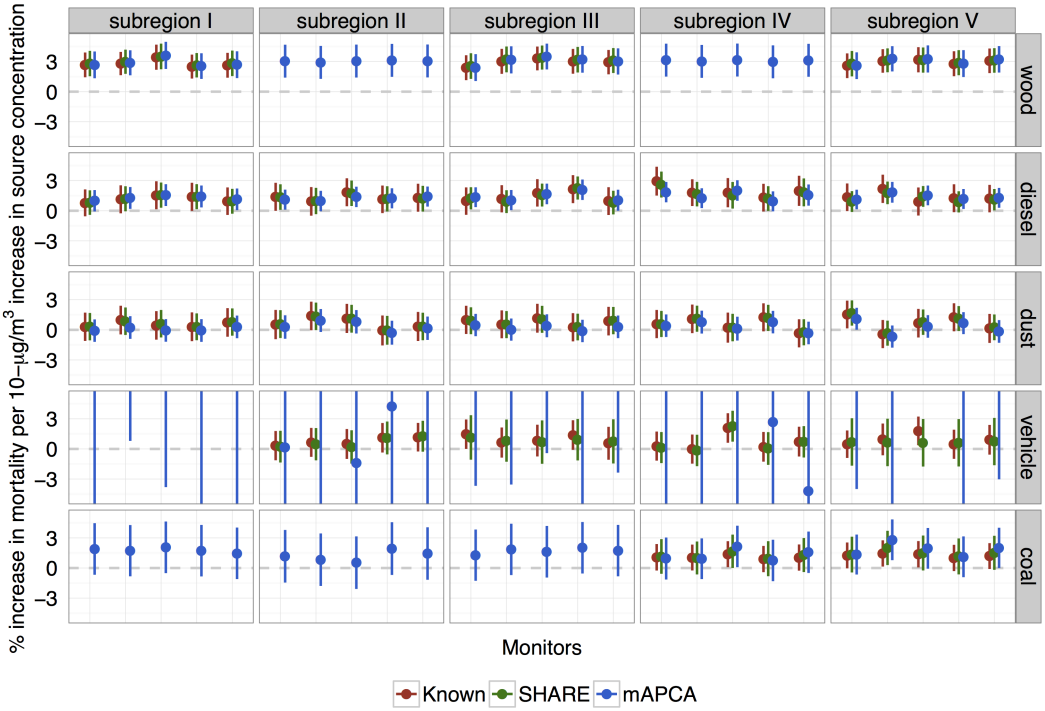

Figure 6.5: Empirical Bayes estimates for each monitor, reported as the percent increase in mortality (95% posterior interval) associated with a 10 $\mu g/m^3$ increase in source concentration under simulation scenario D (25 monitors with 5 monitors in each subregion I-V) using simulated source concentrations (Known), APCA with SHARE (SHARE), and mAPCA.

Table 6.5: Major sources of $PM_{2.5}$ in northeastern US with the number of monitors (out of 41) where the source was identified and the constituents most associated with each source.

Source name	Monitors	Communities	Major constituents
Traffic	33	24	Zinc, manganese, EC, iron, calcium, OC, lead
Soil	33	24	Silicon, aluminum, titanium, calcium, iron
Secondary sulfate	39	23	Sulfate, ammonium, OC, selenium
Fireworks	33	21	Strontium, potassium, copper
Sea salt	27	21	Chlorine, nitrate, sodium ion, bromine
P/V	14	10	Phosphorus, vanadium
Residual oil	7	6	Nickel, vanadium
As/Se/Br	6	5	Arsenic, selenium, bromine

constituents associated with each source, specifically those constituents corresponding to values greater than 0.4 or less than -0.4 in the source signature. When possible, we named sources by matching our sources to sources identified in the literature (Ito et al., 2004). However source names should be interpreted with caution since each identified source may represent any $PM_{2.5}$ source that shares contributing chemical constituents. Figure 6.6 shows the monitors where each major $PM_{2.5}$ source was found (closed circles) and the monitors where the source was not found (plus signs).

We performed a validation substudy of SHARE using our northeastern US dataset. For 10 randomly selected monitors, two researchers independently determined which

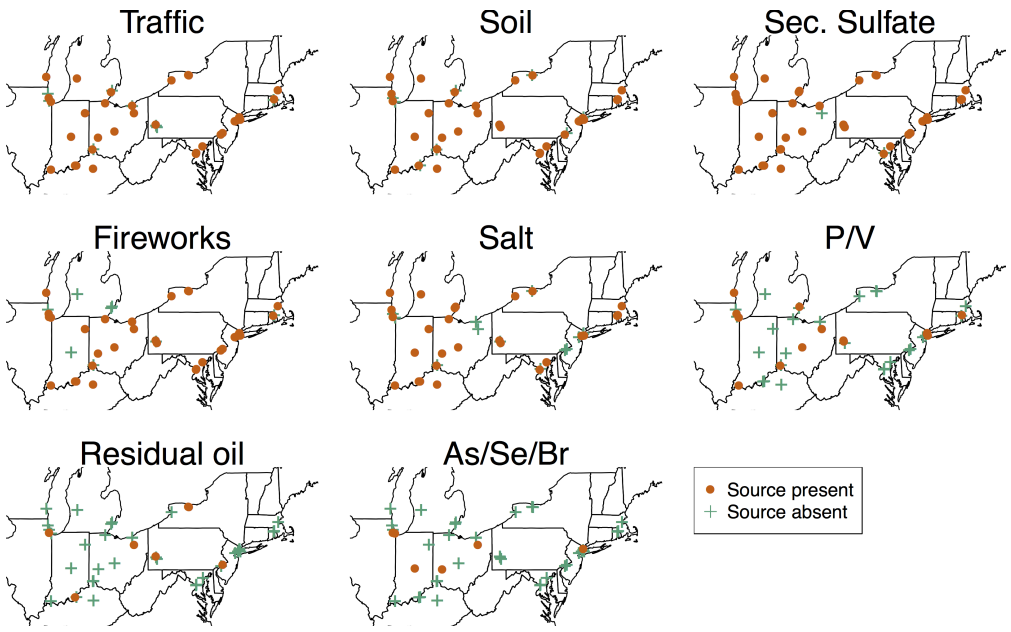

Figure 6.6: Maps corresponding to the 8 regional sources identified in the northeastern US. Each map shows the monitors where that source was found (closed circles) and the monitors where the source was not found (plus signs).

major $PM_{2.5}$ sources were present at each monitor by hand. This method of manually matching sources between monitors is commonly applied in the literature (Ito et al., 2004). We combined monitor source assignments between the two researchers by keeping assignments where the researchers agreed and eliminating assignments where the researchers disagreed. For example, if one researcher called a source "wood" and another called the source "vehicle," we determined that the source did not resemble a major source. We compared results from the manual source identification to source identification using SHARE. The two had good agreement: of 65 sources identified across 10 monitors, SHARE matched the manual approach for 50 sources (76.9%). In another 10 sources, SHARE yielded no assignment because the angle between the monitor-specific source and its closest major $PM_{2.5}$ source exceeded the cutoff of 45 degrees, indicating a poor match. The manual approach does not inherently allow for thresholding poor matches. Excluding cases when matches exceeded the threshold of 45 degrees, SHARE matched the manual approach in 50 of 55 sources across all monitors (90.9%).

In a sensitivity analysis, we also applied SHARE using thresholds larger than 45 degrees and found sources, particularly the P/V, residual oil, and As/Se/Br sources, were identified at more monitors. The other 5 sources were identified at a similar number of monitors regardless of the threshold. To test whether the number of total monitors affects the performance of SHARE, we also identified major sources for the 5 monitors in New York City using successively larger datasets. We applied SHARE using data from (1) only the 5 monitors in New York City (2) communities along the east coast: New York City, NY; Philadelphia, PA; Boston, MA; Providence, RI; Washington, DC; Baltimore, MD and (3) all 41 monitors in the northeast. For the 3 datasets, we manually compared the sources estimated at the 5 New York City

monitors and did not find that the sources identified in the 5 New York City monitors varied substantially between datasets.

6.5.2 Associations between all-cause mortality and $PM_{2.5}$ sources

Our combined mortality and $PM_{2.5}$ constituent dataset had 41 $PM_{2.5}$ speciation monitors located within 24 communities. While most communities (n=12) had only one speciation monitor, the other 12 communities had 2 monitors (n=6), 3 monitors (Pittsburgh, Cleveland, Boston), 4 monitors (Philadelphia), 5 monitors (Chicago), and 7 monitors (New York City). We applied APCA and mAPCA to $PM_{2.5}$ constituent data to estimate source concentrations at each monitor. For communities with multiple monitors, we averaged estimated source concentrations for each day, as is commonly done in similar studies of PM and health (Krall et al., 2013; Peng et al., 2009). For days with only one monitor recording data, we used estimated source concentrations from that monitor. While averaging community-level source concentrations was straightforward for mAPCA, we used SHARE for APCA to match sources between monitors in the same community. SHARE guides pooling source concentrations across monitors, just as SHARE guides pooling estimated health effects across multiple communities. We matched sources between SHARE and mAPCA using the Hungarian method as described in Section 6.3.1. Using mAPCA we did not find a P/V source, but otherwise found the other 7 of 8 sources identified by SHARE (Table 6.5).

We estimated community-level associations between mortality and short-term exposure to $PM_{2.5}$ sources using overdispersed Poisson time series regression models (equation 6.3). Covariates in the model included indicators for day of week and age category ($\leq$64, 65-74, $\geq$75). In addition, to control for confounding by weather, we

included smooth functions (natural spline) of temperature and one-day lag of temperature, each with 3 degrees of freedom. To account for long-term trends in mortality, we included a smooth function of time with 8 degrees of freedom per year. These covariates have been previously used in studies estimating health effects of $PM_{2.5}$ total mass and $PM_{2.5}$ constituents (Krall et al., 2013; Zanobetti and Schwartz, 2009). As in previous studies, we estimated associations between mortality and $PM_{2.5}$ sources for same-day exposure (lag 0), previous-day exposure (lag 1), and exposure 2 days before (lag 2) (Krall et al., 2013; Zanobetti and Schwartz, 2009).

We estimated associations with mortality for the 5 major $PM_{2.5}$ sources identified by SHARE that matched known sources in the northeastern US: traffic, soil, secondary sulfate, sea salt, and residual oil (Ito et al., 2004; Nikolov et al., 2007; Thurston et al., 2011; Hopke et al., 2006). It is common in source apportionment analyses to focus on estimated sources that match known sources of pollution in the area (Ito et al., 2004). For each of the sources, we pooled relevant community-specific associations using a two-level Bayesian hierarchical model. As in the simulation study, APCA results were pooled using SHARE, while mAPCA assumes all sources are the same across monitors. We reported estimated associations as the percent increase in mortality associated with a 10-$\mu g/m^3$ increase in each major $PM_{2.5}$ source. The associations and 95% posterior intervals for exposure lags 0-2 are shown in Figure 6.7. We did not find evidence that $PM_{2.5}$ sources were associated with mortality at any lag, though estimated associations had large standard errors across all sources.

Figure 6.8 shows empirical Bayes estimated mortality risks for 24 communities corresponding to previous-day (Lag 1) exposure to salt and traffic sources of $PM_{2.5}$. Previous studies have shown that associations between $PM_{2.5}$ and mortality

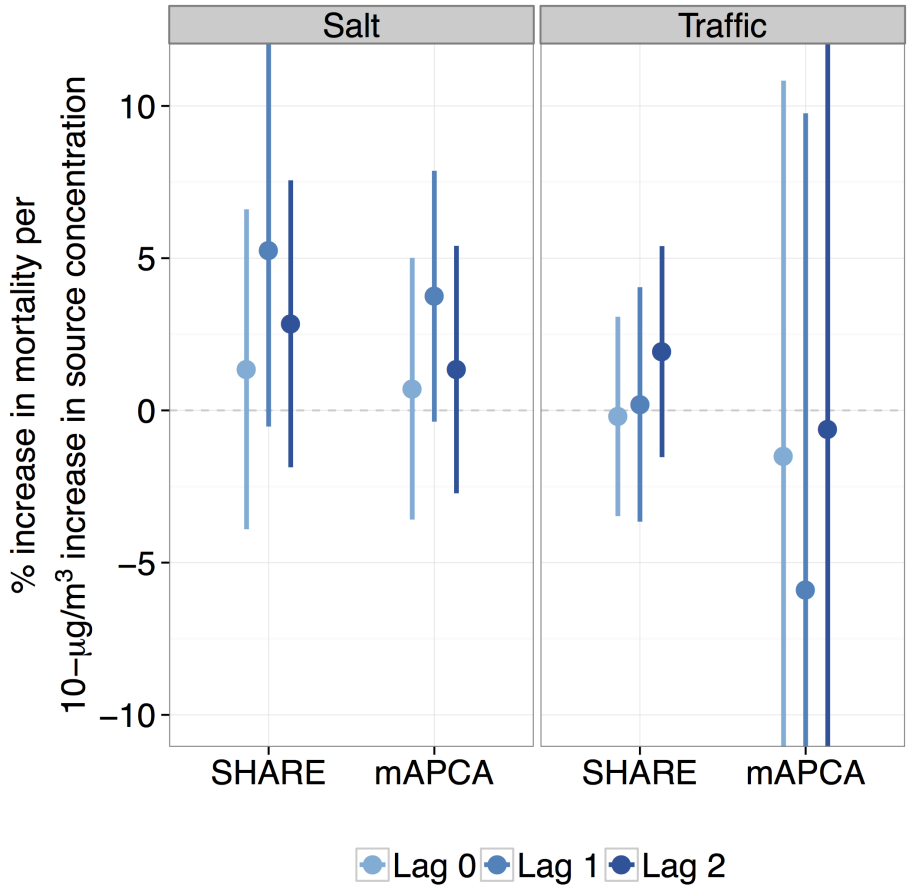

Figure 6.7: Regional percent increase in mortality (95% posterior intervals) associated with a 10-$\mu g/m^3$ increase in same-day (lag 0) previous-day (lag 1), and two days before (lag 2) source concentration for 5 sources identified in the northeastern US. Results are shown for APCA with SHARE (SHARE) and mAPCA.

are strongest at a one-day lag (Krall et al., 2013; Peng et al., 2009), and we did not find evidence of community-specific associations at lags 0 or 2. Two communities (Louisville, KY and Washington, DC) had positive and statistically significant associations between previous-day salt and mortality using both SHARE and mAPCA. However, the point estimates were larger than PM-related health effects reported in the literature (Krall et al., 2013; Mar et al., 2006; Zanobetti and Schwartz, 2009) and likely do not reflect true variability in community-specific associations. SHARE did not identify a salt source in three communities (Detroit, MI; Philadelphia, PA; Providence, RI), and estimated mortality effects are only shown for mAPCA for these communities. We found some evidence of an association between previous-day exposure to traffic and mortality in New York City, NY using SHARE, but little evidence of associations for other communities. All community-specific associations with traffic for mAPCA had large standard errors. The associations for secondary sulfate (not shown) were very similar between APCA with SHARE and mAPCA. For soil, residual oil, and fireworks, 95% posterior intervals were large for both SHARE and mAPCA (results not shown).

6.6 Discussion

Studies of the associations between $PM_{2.5}$ sources and health effects have been generally limited to studies of single ambient monitors. National-level studies of $PM_{2.5}$ and $PM_{2.5}$ constituents more precisely estimate health effects relative to studies of smaller regions (Krall et al., 2013; Zanobetti and Schwartz, 2009). However, source apportionment results cannot easily be pooled across multiple monitors, limiting our ability to estimate regional associations between $PM_{2.5}$ sources and adverse health outcomes. We developed SHARE, a quantitative, reproducible approach to estimate

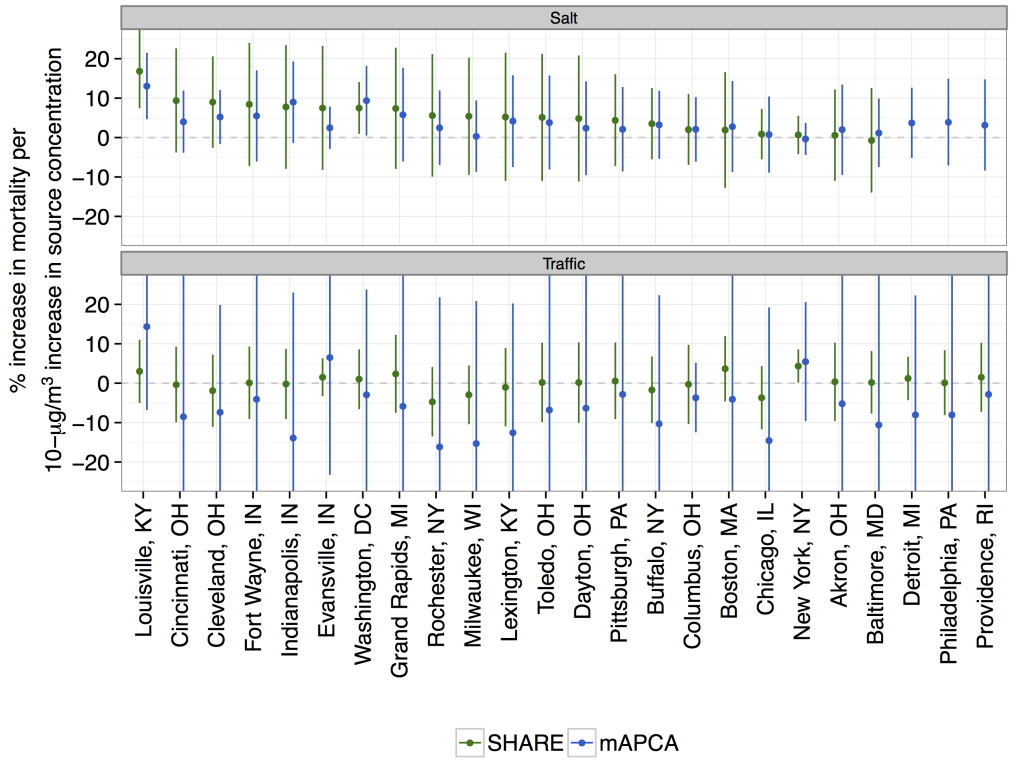

Figure 6.8: Empirical Bayes estimates for each community, reported as the percent increase in mortality (95% posterior interval) associated with a 10 $\mu g/m^3$ increase in previous day (lag 1) salt and traffic $PM_{2.5}$ sources. Results are shown for APCA with SHARE (SHARE) and mAPCA.

major $PM_{2.5}$ sources that are shared across multiple monitors.

We demonstrated that SHARE can be used to identify major $PM_{2.5}$ sources in both a simulation study and in a study of $PM_{2.5}$ sources in the northeastern US. Commonly in practice, sources are matched between monitors using ad hoc, manual interpretations of source apportionment results. We validated our application of SHARE in the northeastern US using evaluations from two independent researchers. SHARE matched the manual approach for 76.9% of sources, but the researchers only agreed with each other for 81.5% of sources. Using SHARE, we were able to estimate the locations of major $PM_{2.5}$ sources across monitors, and found that most sources identified were more regional in nature, such as traffic, soil, secondary sulfate, fireworks, and salt. A previous study found that source chemical compositions varied between monitors in New York City, NY (Ito et al., 2004), and another study found correlations between the monitors varied by source in Atlanta, GA (Marmur et al., 2006). Because of this observed spatial variation in sources within single communities, SHARE might be necessary to estimate sources even within one community. SHARE can also be used along with APCA, a commonly applied source apportionment method (Ito et al., 2004; Hopke et al., 2006), and time series health effects regression models (Mar et al., 2006; Ito et al., 2006) to estimate regional health effects of $PM_{2.5}$ sources. Using SHARE, we estimated the first regional associations between short-term exposure to major $PM_{2.5}$ sources and mortality.

While a comparison between APCA and mAPCA was not the primary focus of this paper, SHARE allowed us to compare the two approaches in estimating regional health effects of $PM_{2.5}$ sources. The mAPCA method assumes sources are the same

across monitors, which is likely an incorrect assumption for multiple monitors spanning large regions. In simulation scenario (B) where sources vary between 5 monitors, the estimated health effects using mAPCA for wood and coal both have larger standard errors compared with APCA using SHARE (Figure 6.4). In this scenario mAPCA estimates mortality effects for wood and coal at all monitors, though these sources are not present at all monitors. Both APCA using SHARE and mAPCA estimate health effects for vehicle with very little precision. Profiles used to generate data for diesel exhaust and motor vehicles shared several constituents including OC and EC (Figure 6.3). Source apportionment methods may not be able to differentiate sources with similar chemical compositions (Marmur et al., 2006). APCA and mAPCA rely on PCA to estimate sources and after estimating wood, diesel, dust, and coal, there is little variation in the data left to estimate vehicle.

When estimating community-specific mortality risk estimates, SHARE can guide where we should estimate associations for a particular source. In our simulation study, SHARE determined that wood was not present in some communities and therefore we did not estimate associations using APCA there. However we found that mAPCA, which assumes sources do not vary spatially, estimated positive, statistically significant associations between wood and mortality in communities where wood burning was not present. Because wood was not present at these communities, the corresponding empirical Bayes estimates were close to the regional estimate, which was positive and statistically significant. The major $PM_{2.5}$ sources identified by SHARE included a P/V source that was not identified by mAPCA, which is likely because SHARE selects major $PM_{2.5}$ sources by weighting sources estimated at all monitors equally. In contrast, mAPCA concatenates concentrations across all monitors and the major sources are more influenced by monitors with more data. In the

northeast, major $PM_{2.5}$ sources were generally present at all monitors. Therefore we did not see substantial differences in estimated community-specific associations between SHARE and mAPCA.

Most source apportionment models, including both mAPCA and APCA, do not yield estimates of uncertainty for the estimated source concentrations. To estimate associations between $PM_{2.5}$ sources and mortality, we treated estimated source concentrations as known in time series regression models and have likely underestimated the uncertainty in estimated health effects. One option for incorporating uncertainty estimates in APCA results is to create bootstrapped confidence intervals of the PCs used to estimate sources (e.g. Babamoradi et al. (2013)). However, bootstrapping confidence intervals for the study of health effects of $PM_{2.5}$ sources is challenging because the data are time series and the outcomes of interest are generally daily counts of mortality or morbidity. Fully Bayesian models can simultaneously estimate $PM_{2.5}$ sources and associations between sources and adverse health outcomes (Nikolov et al., 2007, 2008) and therefore can incorporate uncertainty from unknown source concentrations in estimated health effects of $PM_{2.5}$ sources. Bayesian models have not yet been applied to data from multiple monitors.

SHARE only matches source signatures at a monitor to major $PM_{2.5}$ sources when the angle between sources is less than a pre-specified threshold. We did not find that the sources identified at monitors in the northeastern US were sensitive to the choice of threshold. The best choice of threshold may vary between applications because smaller thresholds decrease misidentification of sources. In contrast, larger thresholds increase power because sources will be identified at more monitors and we can therefore pool information across more monitors. The angles between source signatures at a monitor and major $PM_{2.5}$ sources generally represent the goodness of

matches with smaller angles indicating better matches. When combining source information across multiple monitors, we could incorporate these angles from SHARE to more heavily weight sources closest to major $PM_{2.5}$ sources.

We did not compare APCA with SHARE and mAPCA results to results from other source apportionment models such as PMF and Unmix. SHARE requires a source apportionment model that explicitly uses PCA to estimate sources (e.g. APCA) and therefore cannot be applied in its current form to PMF or Unmix results. In Bayesian source apportionment models, SHARE could be applied to develop priors for whether sources are shared between multiple monitors.

SHARE allows sources to vary between monitors, but it does not use spatial information to determine whether sources are shared across multiple monitors. We would expect that sources would be more similar between neighboring monitors compared with monitors separated by large distances. Jun and Park (2013) estimated the spatial correlation of volatile organic compound sources across 8 sites in Harris County, TX, but they did not allow the chemical composition of sources to vary spatially. SHARE estimates major $PM_{2.5}$ sources by weighting source signatures from all monitors equally, regardless of the amount of data available at each monitor. Another approach would be to estimate major sources by weighting source signatures from each monitor by the amount of available data. However, weighting source signatures in this way may result in only estimating sources present at monitors with many days of data and failing to identify major $PM_{2.5}$ sources that are shared across all monitors.

We demonstrated that SHARE, a novel approach for matching $PM_{2.5}$ sources across multiple monitors, can be applied to estimate major sources of $PM_{2.5}$ across a region. Additionally, SHARE can be used to guide pooling information about sources, including community-specific health effects, across a region. We used SHARE

to estimate regional associations between major $PM_{2.5}$ sources and mortality in the northeastern US.

Chapter 7

Conclusions

Previous studies have found positive associations between short-term exposure to PM air pollution and adverse health outcomes, however these associations could be mediated by one or more properties of PM, including its size, composition, shape, or particle number. Current epidemiologic and toxicological research has found smaller particles, such as $PM_{2.5}$, are more harmful to human health than larger particles (Environmental Protection Agency, 2009). Because $PM_{2.5}$ is a heterogeneous mixture of different chemical constituents, characterizing the toxicity of $PM_{2.5}$ of varying compositions may elucidate which portions of $PM_{2.5}$ are most toxic. This thesis contributes to the growing body of work that has found $PM_{2.5}$ toxicity depends on its composition. In Chapters 3 and 4, we found evidence that EC, OCM, silicon, and sodium ion constituents of $PM_{2.5}$ were more associated with mortality than ammonium, sulfate, and nitrate. A previous large-scale study of $PM_{2.5}$ constituents found evidence of associations between emergency hospitalizations and exposure to OCM and EC, but did not find evidence of associations for other $PM_{2.5}$ constituents (Peng et al., 2009). In general, there is more evidence from both epidemiologic and toxicological studies supporting the toxicity of EC and OCM compared with other

constituents (Rohr and Wyzga, 2012). Ammonium, sulfate, and nitrate are primarily secondary pollutants and have not been found to be harmful in most epidemiologic studies (Schlesinger, 2007). We also estimated associations between major $PM_{2.5}$ sources and mortality in the northeastern US, but did not find evidence that associations were stronger for sources containing OCM or EC. One limitation of this study was that we only estimated associations between mortality and $PM_{2.5}$ sources for 24 communities, compared with 72 communities used to estimate associations between mortality and $PM_{2.5}$ constituents in Chapters 3 and 4. Because sources of $PM_{2.5}$ must be estimated from $PM_{2.5}$ constituent concentrations, associations between $PM_{2.5}$ sources and mortality are difficult to estimate precisely.

Other attributes of $PM_{2.5}$ besides its size and chemical composition could be important for understanding its toxicity. While we focused on estimating health effects of $PM_{2.5}$ sources because of their varying chemical compositions, $PM_{2.5}$ sources may also vary in toxicity because of the age of the aerosol (i.e. fresh or stale) and whether the particle is a fiber. However, data on these attributes of $PM_{2.5}$ sources are not generally available and they have not previously been the focus of epidemiologic studies. Another important attribute of $PM_{2.5}$ could be the particle number of ultrafine PM (PM$<0.1\ \mu m$).

Associations between $PM_{2.5}$ composition and adverse health outcomes can be estimated using different approaches than those discussed in this work. Instead of estimating $PM_{2.5}$ sources from ambient speciation monitors, chemistry transport models are applied to model the formation of secondary particles and transport of pollutants from $PM_{2.5}$ sources (Environmental Protection Agency, 2009). Air quality models use emissions inventories, meteorological data, chemistry transport models, and other data to model pollution (Özkaynak et al., 2013). One example of an air quality

model is the Community Multi-scale Air Quality (CMAQ) model, which is currently used by the US EPA to track pollution throughout the US. Recent models have also incorporated monitoring data and air quality models in a hybrid modeling approach to estimate pollution (Berrocal et al., 2012; Chang et al., 2013). These models have the advantage of using more information to generate more spatially and temporally resolved pollution estimates than are generally available from monitoring data alone. While these alternative models can also be used to estimate associations between $PM_{2.5}$ constituents and adverse health outcomes, many researchers still use monitoring data from networks such as the EPA CSN to estimate health effects. Because speciation monitoring data are readily available from the EPA and can be analyzed without much computing power, the development of methods for ambient monitoring data to improve estimated health effects related to $PM_{2.5}$ composition is critical.

Under the current US National Ambient Air Quality Standards, areas not in attainment of the $PM_{2.5}$ standards must reduce their total $PM_{2.5}$ by mass, without regard to $PM_{2.5}$ composition. Some research has found $PM_{2.5}$ from mobile sources (e.g. diesel exhaust) to be more toxic than other sources of $PM_{2.5}$ (Grahame and Schlesinger, 2010), which might indicate reducing emissions from mobile sources will better protect public health than reducing $PM_{2.5}$ from other sources. In our work, we did not find evidence that $PM_{2.5}$ from traffic was more associated with mortality than other $PM_{2.5}$ sources. One challenge in creating more targeted regulation of $PM_{2.5}$ is that $PM_{2.5}$ sources are not directly measured at the national level. Source apportionment models can be used to estimate sources from available $PM_{2.5}$ speciation data observed at ambient monitors, but source apportionment cannot determine the origin of identified $PM_{2.5}$ sources. Data from emissions inventories or air quality models would be necessary to determine the nature and location of $PM_{2.5}$

sources for regulation purposes.

This thesis developed methods for estimating health effects of $PM_{2.5}$ of varying compositions. Chapters 4 and 5 provided guidance on how measurement error, in the form of spatial misalignment and censored data, should be handled in studies of $PM_{2.5}$ composition. Chapter 6 detailed a novel approach for estimating regional and national-level health effects related to $PM_{2.5}$ sources. $PM_{2.5}$ speciation data have only been collected at the national level in the US since 2000. Therefore, only recently has there been enough speciation monitoring data available for large studies of the health effects associated with exposure to $PM_{2.5}$ chemical constituents and $PM_{2.5}$ sources. The current literature does not definitively support the toxicity of $PM_{2.5}$ of a certain composition, and many epidemiologic studies are conflicted in their findings (Rohr and Wyzga, 2012; Grahame and Schlesinger, 2007). More regional and national-level studies estimating health effects related to $PM_{2.5}$ composition need to be conducted to determine which portions of $PM_{2.5}$ are most toxic.

Chapter 8

Bibliography

Aruga, R. (1997). Treatment of responses below the detection limit: some current techniques compared by factor analysis on environmental data. Analytica Chimica Acta, **354**(1-3), 255–262.

Babamoradi, H., van den Berg, F., and Rinnan, Å. (2013). Bootstrap based confidence limits in principal component analysis–a case study. Chemometrics and Intelligent Laboratory Systems, **120**, 97–105.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). Hierarchical Modeling and Analysis for Spatial Data. CRC Press.

Batalha, J. R. F., Saldiva, P. H. N., Clarke, R. W., Coull, B. A., Stearns, R. C., Lawrence, J., Murthy, G. G. K., Koutrakis, P., and Godleski, J. J. (2002). Concentrated ambient air particles induce vasoconstriction of small pulmonary arteries in rats. Environmental Health Perspectives, **110**(12), 1191–1197.

Bell, M. L., Dominici, F., Ebisu, K., Zeger, S. L., and Samet, J. M. (2007). Spatial and temporal variation in $PM_{2.5}$ chemical composition in the United States for health effects studies. Environmental Health Perspectives, **115**(7), 989–995.

Bell, M. L., Ebisu, K., Peng, R. D., Samet, J. M., and Dominici, F. (2009). Hospital admissions and chemical composition of fine particle air pollution. American Journal of Respiratory and Critical Care Medicine, **179**(12), 1115–1120.

Bell, M. L., Belanger, K., Ebisu, K., Gent, J. F., Lee, H. J., Koutrakis, P., and Leaderer, B. P. (2010). Prenatal exposure to fine particulate matter and birth weight: Variations by particulate constituents and sources. Epidemiology, **21**(6), 884–891.

Bell, M. L., Ebisu, K., and Peng, R. D. (2011). Community-level spatial heterogeneity of chemical constituent levels of fine particulates and implications for epidemiological research. Journal of Exposure Science & Environmental Epidemiology, **21**(4), 372–384.

Bell, M. L., Ebisu, K., Leaderer, B. P., Gent, J. F., Lee, H. J., Koutrakis, P., Wang, Y., Dominici, F., and Peng, R. D. (2013). Associations of $PM_{2.5}$ constituents and sources with hospital admissions: Analysis of four counties in Connecticut and Massachusetts (USA) for persons $\geq$ 65 years of age. Environmental Health Perspectives, **122**(2), 138–144.

Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2012). Space-time data fusion under error in computer model output: An application to modeling air quality. Biometrics, **68**(3), 837–848.

Brown, D. M., Wilson, M. R., MacNee, W., Stone, V., and Donaldson, K. (2001). Size-dependent proinflammatory effects of ultrafine polystyrene particles: a role for surface area and oxidative stress in the enhanced activity of ultrafines. Toxicology and Applied Pharmacology, **175**(3), 191–199.

Burnett, R., Brook, J., Dann, T., Delocla, C., Philips, O., Cakmak, S., Vincent, R., Goldberg, M., and Krewski, D. (2000). Association between particulate- and gas-phase components of urban air polluton and daily mortality in eight Canadian

cities. Inhalation Toxicology, **12**(S4), 15–39.

Buzcu-Guven, B., Brown, S. G., Frankel, A., Hafner, H. R., and Roberts, P. T. (2007). Analysis and apportionment of organic carbon and fine particulate matter sources at multiple sites in the midwestern United States. Journal of the Air & Waste Management Association, **57**(5), 606–619.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific Computing, **16**(6), 1190–1208.

Campen, M. J., Nolan, J. P., Schladweiler, M. C. J., Kodavanti, U. P., Costa, D. L., and Watkinson, W. P. (2002). Cardiac and thermoregulatory effects of instilled particulate matter-associated transition metals in healthy and cardiopulmonary-compromised rats. Journal of Toxicology and Environmental Health, Part A, **65**(20), 1615–1631.

Cao, J., Xu, H., Xu, Q., Chen, B., and Kan, H. (2012). Fine particulate matter constituents and cardiopulmonary mortality in a heavily polluted Chinese city. Environmental Health Perspectives, **120**(3), 373–378.

Carlin, B. P. and Louis, T. A. (2009). Bayesian methods for data analysis. CRC Press.

Chang, H. H., Peng, R. D., and Dominici, F. (2011). Estimating the acute health effects of coarse particulate matter accounting for exposure measurement error. Biostatistics, **12**(4), 637–652.

Chang, H. H., Hu, X., and Liu, Y. (2013). Calibrating modis aerosol optical depth for predicting daily $PM_{2.5}$ concentrations via statistical downscaling. Journal of Exposure Science and Environmental Epidemiology.

Chen, H., Quandt, S. A., Grzywacz, J. G., and Arcury, T. A. (2013). A Bayesian

multiple imputation method for handling longitudinal pesticide data with values below the limit of detection. Environmetrics, **24**(2), 132–142.

Choi, J., Fuentes, M., and Reich, B. J. (2009). Spatial–temporal association between fine particulate matter and daily mortality. Computational Statistics & Data Analysis, **53**(8), 2989–3000.

Cifuentes, L. A., Vega, J., Köpfer, K., and Lave, L. B. (2000). Effect of the fine fraction of particulate matter versus the coarse mass and other pollutants on daily mortality in Santiago, Chile. Journal of the Air & Waste Management Association, **50**(8), 1287–1298.

Costa, D. L. and Dreher, K. L. (1997). Bioavailable transition metals in particulate matter mediate cardiopulmonary injury in healthy and compromised animal models. Environmental Health Perspectives, **105**(S5), 1053–1060.

Crainiceanu, C. M., Caffo, B. S., Luo, S., Zipunnikov, V. M., and Punjabi, N. M. (2011). Population value decomposition, a framework for the analysis of image populations. Journal of the American Statistical Association, **106**(495), 775–790.

Daniels, M. J., Dominici, F., and Zeger, S. L. (2004). Understimation of standard errors in multi-site time series studies. Epidemiology, **15**(1), 57–62.

Darrow, L. A., Klein, M., Flanders, W. D., Waller, L. A., Correa, A., Marcus, M., Mulholland, J. A., Russell, A. G., and Tolbert, P. E. (2009). Ambient air pollution and preterm birth: a time-series analysis. Epidemiology, **20**(5), 689–698.

deCastro, B. R., Wang, L., Mihalic, J. N., Breysse, P. N., Geyh, A. S., and Buckley, T. J. (2008). The longitudinal dependence of black carbon concentration on traffic volume in an urban environment. Journal of the Air & Waste Management Association, **58**(7), 928–939.

Dominici, F., Zeger, S. L., and Samet, J. M. (2000). A measurement error model for

time-series studies of air pollution and mortality. Biostatistics, **1**(2), 157–175.

Dominici, F., Peng, R. D., Bell, M. L., Pham, L., McDermott, A., Zeger, S. L., and Samet, J. M. (2006). Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. Journal of the American Medical Association, **295**(10), 1127–1134.

Dominici, F., Peng, R. D., Zeger, S. L., White, R. H., and Samet, J. M. (2007). Particulate air pollution and mortality in the United States: did the risks change from 1987 to 2000? American Journal of Epidemiology, **166**(8), 880–888.

EarthInfo Inc. (2006). NCDC Summary of the Day. Available: `http://www.earthinfo.com/databases/sd.htm` [Accessed 7 May 2009].

Environmental Protection Agency (1999). Particulate matter ($PM_{2.5}$) Speciation Guidance Document.

Environmental Protection Agency (2009). Integrated Science Assessment for particulate matter.

Everson, P. J. and Morris, C. N. (2000). Inference for multivariate normal hierarchical models. Journal of the Royal Statistical Society, Series B, **62**(2), 399–412.

Farnham, I. M., Singh, A. K., Stetzenbach, K. J., and Johannesson, K. H. (2002). Treatment of nondetects in multivariate analysis of groundwater geochemistry data. Chemometrics and Intelligent Laboratory Systems, **60**(1–2), 265–281.

Francis, R. A., Small, M. J., and VanBriesen, J. M. (2009). Multivariate distributions of disinfection by-products in chlorinated drinking water. Water Research, **43**(14), 3453–3468.

Franklin, M., Zeka, A., and Schwartz, J. (2007). Association between $PM_{2.5}$ and all-cause and specific-cause mortality in 27 US communities. Journal of Exposure Science and Environmental Epidemiology, **17**(3), 279–287.

Franklin, M., Koutrakis, P., and Schwartz, J. (2008). The role of particle composition on the association between $PM_{2.5}$ and mortality. Epidemiology, **19**(5), 680–9.

Fuentes, M. and Smith, R. L. (2001). A new class of nonstationary spatial models. Technical report, North Carolina State University, Department of Statistics.

Fung, K. Y. and Krewski, D. (1999). On measurement error adjustment methods in Poisson regression. Environmetrics, **10**(2), 213–224.

Ganser, G. H. and Hewett, P. (2010). An accurate substitution method for analyzing censored data. Journal of Occupational and Environmental Hygiene, **7**(4), 233–244.

Godleski, J. J., Clarke, R. W., Coull, B. A., Saldiva, P. H. N., Jiang, N.-F., Lawrence, J., and Koutrakis, P. (2002). Composition of inhaled urban air particles determines acute pulmonary responses. Annals of Occupational Hygiene, **46**(s1), 419–424.

Gojova, A., Guo, B., Kota, R. S., Rutledge, J. C., Kennedy, I. M., and Barakat, A. I. (2007). Induction of inflammation in vascular endothelial cells by metal oxide nanoparticles: effect of particle composition. Environmental Health Perspectives, **115**(3), 403–409.

Grahame, T. J. and Schlesinger, R. B. (2007). Health effects of airborne particulate matter: do we know enough to consider regulating specific particle types or sources? Inhalation toxicology, **19**(6-7), 457–481.

Grahame, T. J. and Schlesinger, R. B. (2010). Cardiovascular health and particulate vehicular emissions: a critical evaluation of the evidence. Air Quality, Atmosphere & Health, **3**(1), 3–27.

Gryparis, A., Paciorek, C. J., Zeka, A., Schwartz, J., and Coull, B. A. (2009). Measurement error caused by spatial misalignment in environmental epidemiology. Biostatistics, **10**(2), 258–274.

Guttman, L. (1954). Some necessary conditions for common-factor analysis. Psychometrika, **19**(2), 149–161.

Han, I., Ramos-Bonilla, J. P., Rule, A. M., Mihalic, J. N., Polyak, L. M., Breysse, P. N., and Geyh, A. S. (2011). Comparison of spatial and temporal variations in p-PAH, BC, and p-PAH/BC ratio in six US counties. Atmospheric Environment, **45**(40), 7644–7652.

Harman, H. H. (1976). Modern factor analysis. University of Chicago Press.

Harris, C. W. and Kaiser, H. F. (1964). Oblique factor analytic solutions by orthogonal transformations. Psychometrika, **29**(4), 347–362.

Helsel, D. (2010). Much ado about next to nothing: incorporating nondetects in science. Annals of Occupational Hygiene, **54**(3), 257–262.

Helsel, D. R. (2005a). More than obvious: better methods for interpreting nondetect data. Environmental Science & Technology, **39**(20), 419A–423A.

Helsel, D. R. (2005b). Nondetects and data analysis. Statistics for censored environmental data. Wiley-Interscience.

Helsel, D. R. (2006). Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it. Chemosphere, **65**(11), 2434–2439.

Henry, R. C. (1997). History and fundamentals of multivariate air quality receptor models. Chemometrics and Intelligent Laboratory Systems, **37**(1), 37–42.

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the North Atlantic Ocean. Environmental and Ecological Statistics, **5**(2), 173–190.

Hopke, P. K., Liu, C., and Rubin, D. B. (2001). Multiple imputation for multivariate data with missing and below-threshold measurements: time-series concentrations of pollutants in the arctic. Biometrics, **57**(1), 22–33.

Hopke, P. K., Ito, K., Mar, T., Christensen, W. F., Eatough, D. J., Henry, R. C., Kim,

E., Laden, F., Lall, R., Larson, T. V., et al. (2006). PM source apportionment and health effects: 1. Intercomparison of source apportionment results. Journal of Exposure Science and Environmental Epidemiology, **16**(3), 275–286.

Huang, W., Cao, J., Tao, Y., Dai, L., Lu, S.-E., Hou, B., Wang, Z., and Zhu, T. (2012). Seasonal variation of chemical species associated with short-term mortality effects of $PM_{2.5}$ in Xi'an, a central city in China. American Journal of Epidemiology, **175**(6), 556–566.

Huang, Y.-C. T., Ghio, A. J., Stonehuerner, J., McGee, J., Carter, J. D., Grambow, S. C., and Devlin, R. B. (2003). The role of soluble components in ambient fine particles-induced changes in human lungs and blood. Inhalation Toxicology, **15**(4), 327–342.

Hwang, I. and Hopke, P. K. (2007). Estimation of source apportionment and potential source locations of $PM_{2.5}$ at a west coastal IMPROVE site. Atmospheric Environment, **41**(3), 506–518.

Ito, K., Xue, N., and Thurston, G. (2004). Spatial variation of $PM_{2.5}$ chemical species and source-apportioned mass concentrations in New York City. Atmospheric Environment, **38**(31), 5269–5282.

Ito, K., Christensen, W. F., Eatough, D. J., Henry, R. C., Kim, E., Laden, F., Lall, R., Larson, T. V., Neas, L., Hopke, P. K., and Thurston, G. D. (2006). PM source apportionment and health effects: 2. An investigation of intermethod variability in associations between source-apportioned fine particle mass and daily mortality in Washington, DC. Journal of Exposure Science and Environmental Epidemiology, **16**(4), 300–310.

Ito, K., Mathes, R., Ross, Z., Nadas, A., Thurston, G., and Matte, T. (2011). Fine particulate matter constituents associated with cardiovascular hospitalizations and

mortality in New York City. Environmental Health Perspectives, **119**(4), 467–473.

Jun, M. and Park, E. S. (2013). Multivariate receptor models for spatially correlated multipollutant data. Technometrics, **55**(3), 309–320.

Kamakura, W. A. and Wedel, M. (2001). Exploratory Tobit factor analysis for multivariate censored data. Multivariate Behavioral Research, **36**(1), 53–82.

Kavouras, I. G., Koutrakis, P., Cereceda-Balic, F., and Oyola, P. (2001). Source apportionment of PM_{10} and $PM_{2.5}$ in five Chilean cities using factor analysis. Journal of the Air & Waste Management Association, **51**(3), 451–464.

Kim, E., Hopke, P. K., Paatero, P., and Edgerton, E. S. (2003). Incorporation of parametric factors into multilinear receptor model studies of Atlanta aerosol. Atmospheric Environment, **37**(36), 5009–5021.

Kim, S.-Y., Peel, J. L., Hannigan, M. P., Dutton, S. J., Sheppard, L., Clark, M. L., and Vedal, S. (2012). The temporal lag structure of short-term associations of fine particulate matter chemical constituents and cardiovascular and respiratory hospitalizations. Environmental Health Perspectives, **120**(8), 1094–1099.

Kleinman, M. T., Sioutas, C., Froines, J. R., Fanning, E., Hamade, A., Mendez, L., Meacher, D., and Oldham, M. (2007). Inhalation of concentrated ambient particulate matter near a heavily trafficked road stimulates antigen-induced airway responses in mice. Inhalation Toxicology, **19**(s1), 117–126.

Krall, J. R., Anderson, G. B., Dominici, F., Bell, M. L., and Peng, R. D. (2013). Short-term exposure to particulate matter constituents and mortality in a national study of US urban communities. Environmental Health Perspectives, **121**(10), 1148–1153.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. Naval Research Logistics Quarterly, **2**(1-2), 83–97.

Laden, F., Neas, L. M., Dockery, D. W., and Schwartz, J. (2000). Association of fine particulate matter from different sources with daily mortality in six U.S. cities. Environmental Health Perspectives, **108**(10), 941–947.

Larson, T., Gould, T., Simpson, C., Liu, L.-J. S., Claiborn, C., and Lewtas, J. (2004). Source apportionment of indoor, outdoor, and personal $PM_{2.5}$ in Seattle, Washington, using Positive Matrix Factorization. Journal of the Air & Waste Management Association, **54**(9), 1175–1187.

Lee, D., Ferguson, C., and Mitchell, R. (2009). Air pollution and health in Scotland: a multicity study. Biostatistics, **10**(3), 409–423.

Lee, S., Liu, W., Wang, Y., Russell, A. G., and Edgerton, E. S. (2008). Source apportionment of $PM_{2.5}$: Comparing PMF and CMB results for four ambient monitoring sites in the southeastern United States. Atmospheric Environment, **42**(18), 4126–4137.

Levy, J. I., Diez, D., Dou, Y., Barr, C. D., and Dominici, F. (2012). A meta-analysis and multisite time-series analysis of the differential toxicity of major fine particulate matter constituents. American Journal of Epidemiology, **175**(11), 1091–1099.

Lingwall, J. W., Christensen, W. F., and Reese, C. S. (2008). Dirichlet based Bayesian multivariate receptor modeling. Environmetrics, **19**(6), 618–629.

Lippmann, M., Ito, K., Hwang, J.-S., Maciejczyk, P., and Chen, L.-C. (2006). Cardiovascular effects of nickel in ambient air. Environmental Health Perspectives, **114**(11), 1662–1669.

Luna, X. D. and Genton, M. G. (2005). Predictive spatio-temporal models for spatially sparse environmental data. Statistica Sinica, **15**(2), 547–568.

Mar, T. F., Norris, G. A., Koenig, J. Q., and Larson, T. V. (2000). Associations between air pollution and mortality in Phoenix, 1995-1997. Environmental Health

Perspectives, **108**(4), 347–353.

Mar, T. F., Ito, K., Koenig, J. Q., Larson, T. V., Eatough, D. J., Henry, R. C., Kim, E., Laden, F., Lall, R., Neas, L., et al. (2006). PM source apportionment and health effects. 3. Investigation of inter-method variations in associations between estimated source contributions of $PM_{2.5}$ and daily mortality in Phoenix, AZ. Journal of Exposure Science and Environmental Epidemiology, **16**(4), 311–320.

Marmur, A., Unal, A., Mulholland, J. A., and Russell, A. G. (2005). Optimization-based source apportionment of $PM_{2.5}$ incorporating gas-to-particle ratios. Environmental Science & Technology, **39**(9), 3245–3254.

Marmur, A., Park, S.-K., Mulholland, J. A., Tolbert, P. E., and Russell, A. G. (2006). Source apportionment of $PM_{2.5}$ in the southeastern united states using receptor and emissions-based models: Conceptual differences and implications for time-series health studies. Atmospheric Environment, **40**(14), 2533–2551.

Maykut, N. N., Lewtas, J., Kim, E., and Larson, T. V. (2003). Source apportionment of $PM_{2.5}$ at an urban IMPROVE site in Seattle, Washington. Environmental Science & Technology, **37**(22), 5135–5142.

McDonald, J. D., Zielinska, B., Sagebiel, J. C., McDaniel, M. R., and Mousset-Jones, P. (2003). Source apportionment of airborne fine particulate matter in an underground mine. Journal of the Air & Waste Management Association, **53**(4), 386–395.

Mostofsky, E., Schwartz, J., Coull, B. A., Koutrakis, P., Wellenius, G. A., Suh, H. H., Gold, D. R., and Mittleman, M. A. (2012). Modeling the association between particle constituents of air pollution and health outcomes. American Journal of Epidemiology, **176**(4), 317–326.

Muthén, B. O. (1989). Tobit factor analysis. British Journal of Mathematical and

Statistical Psychology, **42**(2), 241–250.

National Research Council (2004). Research Priorities for Airborne Particulate Matter: IV. Continuing Research Progress. National Research Council of the National Academies.

Nel, A. (2005). Air pollution-related illness: effects of particles. Science, **308**(5723), 804–806.

Nikolov, M. C., Coull, B. A., Catalano, P. J., and Godleski, J. J. (2007). An informative Bayesian structural equation model to assess source-specific health effects of air pollution. Biostatistics, **8**(3), 609–624.

Nikolov, M. C., Coull, B. A., Catalano, P. J., Diaz, E., and Godleski, J. J. (2008). Statistical methods to evaluate health effects associated with major sources of air pollution: a case-study of breathing patterns during exposure to concentrated boston air particles. Journal of the Royal Statistical Society: Series C (Applied Statistics), **57**(3), 357–378.

Nikolov, M. C., Coull, B. A., Catalano, P. J., and Godleski, J. J. (2011). Multiplicative factor analysis with a latent mixed model structure for air pollution exposure assessment. Environmetrics, **22**(2), 165–178.

Norris, G., Vedantham, R., Duvall, R., and Henry, R. C. (2007). EPA Unmix 6.0 fundamentals & user guide. US Environmental Protection Agency, Washington DC.

Norris, G., Vedantham, R., Wade, K., Brown, S., Prouty, J., and Foley, C. (2008). EPA Positive Matrix Factorization 3.0 fundamentals & user guide. US Environmental Protection Agency, Washington DC.

Ostro, B., Broadwin, R., Green, S., Feng, W.-Y., and Lipsett, M. (2006). Fine particulate air pollution and mortality in nine California counties: results from CALFINE.

Environmental Health Perspectives, **114**(1), 29–33.

Ostro, B., Feng, W.-Y., Broadwin, R., Green, S., and Lipsett, M. (2007). The effects of components of fine particulate air pollution on mortality in California: results from CALFINE. Environmental Health Perspectives, **115**(1), 13–19.

Özkaynak, H., Baxter, L. K., Dionisio, K. L., and Burke, J. (2013). Air pollution exposure prediction approaches used in air pollution epidemiology studies. Journal of Exposure Science and Environmental Epidemiology, **23**(6), 566–572.

Paatero, P. (1999). The multilinear engine: A table-driven, least squares program for solving multilinear problems, including the n-way parallel factor analysis model. Journal of Computational and Graphical Statistics, **8**(4), 854–888.

Paatero, P. and Hopke, P. K. (2003). Discarding or downweighting high-noise variables in factor analytic models. Analytica Chimica Acta, **490**(1–2), 277–289.

Paatero, P. and Tapper, U. (1994). Positive Matrix Factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics, **5**(2), 111–126.

Paciorek, C. J. and Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. Environmetrics, **17**(5), 483–506.

Papadimitriou, C. H. and Steiglitz, K. (1998). Combinatorial optimization: algorithms and complexity. Courier Dover Publications.

Peng, R. D. and Bell, M. L. (2010). Spatial misalignment in time series studies of air pollution and health data. Biostatistics, **11**(4), 720–740.

Peng, R. D., Dominici, F., Pastor-Barriuso, R., and Zeger, S. L. (2005). Seasonal analyses of air pollution and mortality in 100 US cities. American Journal of Epidemiology, **161**(6), 585–594.

Peng, R. D., Chang, H. H., Bell, M. L., McDermott, A., Zeger, S. L., Samet, J. M.,

and Dominici, F. (2008). Coarse particulate matter air pollution and hospital admissions for cardiovascular and respiratory diseases among Medicare patients. Journal of the American Medical Association, **299**(18), 2172–2179.

Peng, R. D., Bell, M. L., Geyh, A. S., McDermott, A., Zeger, S. L., Samet, J. M., and Dominici, F. (2009). Emergency admissions for cardiovascular and respiratory diseases and the chemical composition of fine particle air pollution. Environmental Health Perspectives, **117**(6), 957–963.

Polissar, A. V., Hopke, P. K., and Poirot, R. L. (2001). Atmospheric aerosol over Vermont: chemical composition and sources. Environmental Science & Technology, **35**(23), 4604–4621. PMID: 11770762.

Pope, C. A. and Dockery, D. W. (2006). Health effects of fine particulate air pollution: lines that connect. Journal of the Air and Waste Management Association, **56**(6), 709–742.

Pope, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., and Thurston, G. D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. Journal of the American Medical Association, **287**(9), 1132–1141.

Pope, C. A., Burnett, R. T., Thruston, G. D., Calle, E. E., Thun, M. J., Krewski, D., and Godleski, J. J. (2004). Cardiovascular mortality and long-term exposure to particulate air pollution: Epidemiological evidence of general pathophysiological pathways of disease. Circulation, **109**(1), 71–77.

Puett, R. C., Hart, J. E., Yanosky, J. D., Paciorek, C., Schwartz, J., Suh, H., Speizer, F. E., and Laden, F. (2009). Chronic fine and coarse particulate exposure, mortality, and coronary heart disease in the Nurses Health Study. Environmental Health Perspectives, **117**(11), 1697–1701.

Querol, X., Alastuey, A., Rodriguez, S., Plana, F., Ruiz, C. R., Cots, N., Massagué, G., and Puig, O. (2001). PM_{10} and $PM_{2.5}$ source apportionment in the Barcelona metropolitan area, Catalonia, Spain. Atmospheric Environment, **35**(36), 6407–6419.

R Core Team (2012). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rizzo, M. J. and Scheff, P. A. (2007). Fine particulate source apportionment using data from the USEPA speciation trends network in Chicago, Illinois: Comparison of two source apportionment models. Atmospheric Environment, **41**(29), 6276–6288.

Rohr, A. C. and Wyzga, R. E. (2012). Attributing health effects to individual particulate matter constituents. Atmospheric Environment, **62**(0), 130–152.

Rohr, A. C., Wagner, J. G., Morishita, M., Kamal, A., Keeler, G. J., and Harkema, J. R. (2010). Cardiopulmonary responses in spontaneously hypertensive and Wistar-Kyoto rats exposed to concentrated ambient particles from Detroit, Michigan. Inhalation Toxicology, **22**(6), 522–533.

Saldiva, P. H. N., Clarke, R. W., Coull, B. A., Stearns, R. C., Lawrence, J., Murthy, G. G. K., Diaz, E., Koutrakis, P., Suh, H., Tsuda, A., and Godleski, J. J. (2002). Lung inflammation induced by concentrated ambient air particles is related to particle composition. American Journal of Respiratory and Critical Care Medicine, **165**(12), 1610–1617.

Salnikow, K., Li, X., and Lippmann, M. (2004). Effect of nickel and iron co-exposure on human lung cells. Toxicology and Applied Pharmacology, **196**(2), 258–265.

Samet, J. M., Dominici, F., Curriero, F. C., Coursac, I., and Zeger, S. L. (2000a). Fine particulate air pollution and mortality in 20 US cities, 1987–1994. New England

Journal of Medicine, **343**(24), 1742–1749.

Samet, J. M., Zeger, S. L., Dominici, F., Curriero, F., Coursac, I., Dockery, D. W., Schwartz, J., and Zanobetti, A. (2000b). The National Morbidity, Mortality, and Air Pollution Study, Part II: Morbidity and Mortality from Air Pollution in the United States. Health Effects Institute, Cambridge MA.

Sarnat, J. A., Marmur, A., Klein, M., Kim, E., Russell, A. G., Sarnat, S. E., Mulholland, J. A., Hopke, P. K., and Tolbert, P. E. (2008). Fine particle sources and cardiorespiratory morbidity: an application of chemical mass balance and factor analytical source-apportionment methods. Environmental Health Perspectives, **116**(4), 459–466.

Schlesinger, R. B. (2007). The health impact of common inorganic components of fine particulate matter ($PM_{2.5}$) in ambient air: a critical review. Inhalation Toxicology, **19**(10), 811–832.

Schwarze, P., Øvrevik, J., Låg, M., Refsnes, M., Nafstad, P., Hetland, R., and Dybing, E. (2006). Particulate matter properties and health effects: consistency of epidemiological and toxicological studies. Human and Experimental Toxicology, **25**, 559–579.

Seagrave, J., Knall, C., McDonald, J. D., and Mauderly, J. L. (2004). Diesel particulate material binds and concentrates a proinflammatory cytokine that causes neutrophil migration. Inhalation Toxicology, **16**(s1), 93–98.

Song, X.-H., Polissar, A. V., and Hopke, P. K. (2001). Sources of fine particle composition in the northeastern US. Atmospheric Environment, **35**(31), 5277–5286.

Song, Y., Tang, X., Xie, S., Zhang, Y., Wei, Y., Zhang, M., Zeng, L., and Lu, S. (2007). Source apportionment of $PM_{2.5}$ in Beijing in 2004. Journal of Hazardous Materials, **146**(1–2), 124–130.

Strickland, M. J., Gass, K. M., Goldman, G. T., and Mulholland, J. A. (2013). Effects of ambient air pollution measurement error on health effect estimates in time-series studies: a simulation-based analysis. Journal of Exposure Science and Environmental Epidemiology.

Szpiro, A. A., Paciorek, C. J., and Sheppard, L. (2011). Does more accurate exposure prediction necessarily improve health effect estimates? Epidemiology, **22**(5), 680–685.

Thurston, G. D. and Spengler, J. D. (1985). A quantitative assessment of source contributions to inhalable particulate matter pollution in metropolitan Boston. Atmospheric Environment, **19**(1), 9–25.

Thurston, G. D., Ito, K., and Lall, R. (2011). A source apportionment of U.S. fine particulate matter air pollution. Atmospheric Environment, **45**(24), 3924–3936.

Tolbert, P. E., Klein, M., Peel, J. L., Sarnat, S. E., and Sarnat, J. A. (2007). Multipollutant modeling issues in a study of ambient air quality and emergency department visits in Atlanta. Journal of Exposure Science and Environmental Epidemiology, **17**(S2), S29–35.

Urch, B., Brook, J. R., Wasserstein, D., Brook, R. D., Rajagopalan, S., Corey, P., and Silverman, F. (2004). Relative contributions of $PM_{2.5}$ chemical constituents to acute arterial vasoconstriction in humans. Inhalation Toxicology, **16**(6-7), 345–352.

US Census Bureau (2013). Census 2000 Gateway. Available: `http://www.census.gov/main/www/cen2000.html` [Accessed 20 August 2013].

Veronesi, B., de Haar, C., Lee, L., and Oortgiesen, M. (2002). The surface charge of visible particulate matter predicts biological activation in human bronchial epithelial cells. Toxicology and Applied Pharmacology, **178**(3), 144–154.

Wang, T., Lang, G. D., Moreno-Vinasco, L., Huang, Y., Goonewardena, S. N., Peng, Y.-J., Svensson, E. C., Natarajan, V., Lang, R. M., Linares, J. D., et al. (2012). Particulate matter induces cardiac arrhythmias via dysregulation of carotid body sensitivity and cardiac sodium channels. American Journal of Respiratory Cell and Molecular Biology, **46**(4), 524–531.

Wellenius, G. A., Coull, B. A., Godleski, J. J., Koutrakis, P., Okabe, K., Savage, S. T., Lawrence, J. E., Murthy, G. G. K., and Verrier, R. L. (2003). Inhalation of concentrated ambient air particles exacerbates myocardial ischemia in conscious dogs. Environmental Health Perspectives, **111**(4), 402–408.

Yanosky, J. D., Paciorek, C. J., and Suh, H. H. (2009). Predicting chronic fine and coarse particulate exposures using spatiotemporal models for the northeastern and midwestern united states. Environmental Health Perspectives, **117**(4), 522–529.

Zanobetti, A. and Schwartz, J. (2009). The effect of fine and coarse particulate air pollution on mortality: a national analysis. Environmental Health Perspectives, **117**(6), 898–903.

Zanobetti, A., Franklin, M., Koutrakis, P., and Schwartz, J. (2009). Fine particulate air pollution and its components in association with cause-specific emergency admissions. Environmental Health, **8**(58), 1–12.

Zeger, S. L., Thomas, D., Dominici, F., Samet, J. M., Schwartz, J., Dockery, D., and Cohen, A. (2000). Exposure measurement error in time-series studies of air pollution: Concepts and consequences. Environmental Health Perspectives, **108**(5), 419–426.

Zhao, Y., Usatyuk, P. V., Gorshkova, I. A., He, D., Wang, T., Moreno-Vinasco, L.,

Geyh, A. S., Breysse, P. N., Samet, J. M., Spannhake, E. W., et al. (2009). Regulation of COX-2 expression and IL-6 release by particulate matter in airway epithelial cells. American Journal of Respiratory Cell and Molecular Biology, **40**(1), 19–30.

Zhou, J., Ito, K., Lall, R., Lippmann, M., and Thurston, G. (2011). Time-series analysis of mortality effects of fine particulate matter components in Detroit and Seattle. Environmental Health Perspectives, **119**(4), 461–466.

Zidek, J. V., Wong, H., Le, N. D., and Burnett, R. (1996). Causality, measurement error and multicollinearity in epidemiology. Environmetrics, **7**(4), 441–451.

Jenna R. Krall

Biographical information

Date of birth: January 17, 1986

Place of birth: Pittsburgh, Pennsylvania

Email: jenna.krall@gmail.com

Phone: 412-965-2012

Website: jennakrall.com

Education

Ph.D. Biostatistics, Johns Hopkins Bloomberg School of Public Health, 2014

Thesis title: Statistical methods for linking the chemical composition of particulate matter to health outcomes

Advisor: Roger Peng

Certificate in Gerontology, 2013

B.A. Mathematics and Dance, George Mason University, 2008.

summa cum laude

Awards

2013	Louis I. and Thomas D. Dublin Award for the Advancement of Epidemiology and Biostatistics
2013	Second place, Research on Aging Showcase Poster Competition
2008	Genevieve Feinstein Award in Cryptography
2007,2008	Undergraduate Research Apprenticeship Grant Recipient
2006	Amer Beslagic Award in Mathematics
2004-2008	George Mason University Scholarship

Publications

Published or in press

Krall JR, Anderson GB, Dominici F, Bell ML, Peng RD. 2013. Short-term Exposure to Particulate Matter Constituents and Mortality in a National Study of U.S. Urban Communities, Environmental Health Perspectives, 121(10) 1148-1153.

Anderson GB, **Krall JR**, Peng RD, Bell ML. 2012. Is the Relationship Between Ozone and Mortality Confounded by Chemical Components of Particulate Matter? Analysis of 7 Components in 57 United States Communities, American Journal of Epidemiology, 176(8) 726-732.

Farley JE, Ross T, **Krall JR**, Hayat M, Caston-Gaa A, Perl T, Carroll, KC. 2012. Prevalence, Risk Factors and Molecular Epidemiology of Methicillin Resistant Staphylococcus aureus Nasal and Axillary Colonization among Psychiatric Patients on Admission to The Johns Hopkins Hospital, American Journal of Infection Control, 41(3) 199-203.

Turnbull AE, **Krall JR**, Ruhl PA, Curtis JR, Halpern SD, Lau BM, Needham DM. 2013. A scenario-based, randomized trial of patient wishes and functional prognosis on intensivist intent to discuss withdrawing life support, Critical Care Medicine, to appear

Eisbach SS, Cluxton-Keller F, Harrison J, **Krall JR**, Hayat M, Gross D. 2013. Characteristics of Temper Tantrums in Preschoolers with Disruptive Behavior in a Clinic Setting, Journal of Psychosocial Nursing and Mental Health Services, to appear.

Under revision

Krall JR, Carlson MC, Fried LP, Xue QL. Examining the Dynamic, Bidirectional Associations Between Cognitive and Physical Functions, submitted to American Journal of Epidemiology

In Preparation

Powell H, **Krall JR**, Wang Y, Bell ML, Peng RD. Ambient Coarse Particulate Matter and Hospital Admissions in the Medicare Cohort Air Pollution Study, 1999-2010

Krall JR, Simpson CH, Peng RD. Censoring Adjustment Methods for Source Apportionment Models

Krall JR, Hackstadt AJ, Peng RD. A Method to Identify Regional Particulate Matter Sources and their Health Effects.

Professional Experience

2013	NCAR/CDC Colloquium on Climate and Health
2009-2013	Predoctoral Fellow, Epidemiology and Biostatistics of Aging Training Grant
	Advisors: Qian-Li Xue and Michelle Carlson
2010-2011	Biostatistician, Johns Hopkins University School of Nursing Biostatistics Consulting Service
2008	Data Assistant, The Improving Care of Acute Lung Injury Patients (ICAP) study, Johns Hopkins Hospital
2007-2008	Research Assistant, Statistical Assessment Service (STATS), Washington, D.C.
2006-2008	Research Assistant, Measurement/ Research Methodology/ Evaluation/ Statistics (MRES) Research Lab,
	George Mason University Department of Psychology
2005-2006	Intern, GlaxoSmithKline Pharmaceuticals

Presentations

Talks

2014	"Censoring Adjustment Methods for Source Apportionment Models"
	ENAR Spring Meeting, Baltimore, MD, contributed
2013	"Estimating health effects of particulate matter sources in the presence of censored air pollution concentrations"
	Joint Statistical Meetings, Montréal, Canada, topic contributed
2012	"Assessing the Feed-Forward and Feedback Relationship between Cognitive and Physical Functions"
	Gerontological Society of America Annual Meeting, San Diego, CA symposium
2012	"Mortality Effects of Particulate Matter Constituents in a National Study of U.S.Urban Communities"
	ENAR Spring Meeting, Washington, DC, contributed
2011	"Accounting for Spatial Misalignment in a National Study of $PM_{2.5}$ Constituents and Mortality"
	ENAR Spring Meeting, Miami, FL, contributed

Posters

2013 "The Impact of Values Below the Minimum Detection Limit on Source Apportionment Results"

ENAR Spring Meeting, Orlando, FL

2012 "Mortality Effects of Particulate Matter Constituents in a National Study of U.S. Urban Communities"

International Society for Environmental Epidemiology Conference, Columbia, SC

2007 "Predicting Baseball Winners Using Just Noticeable Differences"

Association for Psychological Science Annual Convention, Washington, DC

Teaching

Lead Teaching Assistant

2011 Lead Teaching Assistant for R Learning, Biostatistics in Public Health, Johns Hopkins University

2010,2013 Statistical Methods in Public Health I-III, JHSPH

Guest Lecturer

2012,2013 Statistical Methods in Public Health I-II, JHSPH

2010,2011 Statistical Literacy and Reasoning in Nursing Research, Johns Hopkins University School of Nursing

Teaching Assistant

2012 Introduction to Statistical Computing, JHSPH

2011 MPH Capstone Project, JHSPH

2011,2012 Advanced Statistical Computing, JHU-Nanjing University Exchange Program, JHSPH

2012 Computing for Data Analysis, Coursera

2011 Survival Analysis, JHU-Nanjing University Exchange Program

2010,2011 Data Analysis Workshop I-II, JHSPH

2010 Statistical Methods in Public Health III-IV, JHSPH

2009 Biostatistics in Public Health, Johns Hopkins University

Service

2011-2012	Student Representative to the Biostatistics faculty
2010-2011	Organizer, Hopkins Biostatistics Journal Club
2009-2010	Organizer, Hopkins Biostatistics Computing Club